INFORMS Journal on Computing
Vol. 24, No. 1, Winter 2012, pp. 10–28
ISSN 1091-9856 (print) | ISSN 1526-5528 (online)

http://dx.doi.org/10.1287/ijoc.1110.0452
© 2012 INFORMS

Transient Queueing Analysis

William H. Kaczynski
Department of Mathematical Sciences, United States Military Academy, West Point,
New York 10996, william.kaczynski@usma.edu

Lawrence M. Leemis, John H. Drew
Department of Mathematics, College of William & Mary, Williamsburg, Virginia 23187
{leemis@math.wm.edu, jhdrew@math.wm.edu}

The exact distribution of the nth customer's sojourn time in an $M/M/s$ queue with k customers initially present is derived. Algorithms for computing the covariance between sojourn times for an $M/M/1$ queue with k customers present at time 0 are also developed. Maple computer code is developed for practical application of transient queue analysis for many system measures of performance without regard to traffic intensity (i.e., the system may be unstable with traffic intensity greater than 1).

Key words: exponential distribution; Poisson process; queueing theory
History: Accepted by Winfried Grassmann, Area Editor for Computational Probability and Analysis; received October 2009; revised May 2010; accepted January 2011. Published online in *Articles in Advance* June 17, 2011.

1. Introduction

Many traditional simulation studies analyze queueing systems in steady state, requiring appropriate warm-up periods and associated long simulation runs. However, in many cases, the system being modeled never reaches steady state; thus steady-state simulation results do not accurately portray the system behavior. The ability to analyze transient results associated with such models is often complicated by intractable theory, leaving simulation as the only method for analysis. Further complicating the transient analysis is the effect of initial conditions (Kelton and Law 1985). Because steady-state results depend on running the system long enough to negate the impact of initial conditions, these steady-state results reveal nothing about the transient behavior of the queueing system. Our purpose here is to combine new and existing results in transient queueing analysis with a symbolic engine in computational probability.

There are many classes of queueing systems in which a transient analysis is required; e.g., service businesses often model queues that never reach equilibrium. The need to develop theory for transient results, as opposed to steady-state results, has resulted in wide literature in this area. Initial work in transient analysis ironically appeared as an attempt to measure when a system achieved equilibrium. Law (1975) notes the consequences of failing to adequately account for the initial transient period, which led Gafarian et al. (1976) to develop a comprehensive framework for the initial transient problem. Morisaku (1976) addresses the time to equilibrium in simulations modeling the $M/M/1$ queue and provides schematics for the transition probabilities given $k \geq 0$ customers initially present at time $t = 0$. Grassmann (1977) compares three methods for finding transient solutions in Markovian queueing systems—Runge–Kutta, Liou's method, and randomization—where randomization is shown as superior for large sparse transition matrices. Pegden and Rosenshine (1982) provide a closed-form expression for the probability of exactly i arrivals and j servicings over a time horizon of length t in an $M/M/1$ queue starting empty and idle; this expression allows the calculation of certain performance measures for a specified time period. Odoni and Roth (1983) take an empirical approach to compare observed and predicted transient-state queue lengths for the $M/M/1$ queue, noting that for small values of t, the expected queue length is strongly influenced by initial conditions, and they provide a good approximation for an upper bound of the time to steady state. Kelton and Law (1985) consider the $M/M/s$ $(s \geq 1)$ queue and provide expressions to calculate the probabilities of having up to $n + k$ customers in the system upon the arrival of the nth customer, where k is the number of customers in the system at time $t = 0$. Kelton and Law then apply these calculations to a variety of measures of performance with implications to convergence on steady-state delays, and they offer methods for choosing queue initialization in simulation. Much of the work in this paper is motivated by their results. Kelton (1985) extends the previous work by considering $M/E_m/1$ and $E_m/M/1$ queues. Parthasarathy (1987) provides a transient solution for the probability that there are n customers in the system at time t for an $M/M/1$ queue. Abate and Whitt (1988) use Laplace transformations to analyze some transient results of interest in the $M/M/1$ queue. Leguesdron et al. (1993) provide transient probabilities for the $M/M/1$ queue by inverting the generating function of the uniformized Markov chain describing the $M/M/1$ process. Transient distributions of cumulative reward are calculated by de Souza e Silva et al. (1995) using uniformization, where the distribution of cumulative reward is over a finite interval with reward rates represented by Markov model states. Grassmann (2008) investigates warm-up periods in simulation and shows that in many cases these warm-up periods should not be used, especially if the simulation begins in a high probability state. In this paper we focus on the transient analysis of the $M/M/1$ and the more general $M/M/s$ queues—specifically, on the distribution of the nth customer's sojourn time, which is the sum of the nth customer's delay and service times. Almost all the above-mentioned references address measures of performance specified over a finite time interval and are the results of numerical work. This is in stark contrast to the measures proposed here, which are based on the exact distribution of specific customers. Rather than arriving at the results numerically, a computer algebra system is utilized that offers exact measures of performance based on a given number of customers.

The $M/M/s$ queue is defined in §2 for a positive integer s, and a method is given for calculating the probability distribution of the number of customers an arriving customer sees upon arrival to an $M/M/s$ queue. Section 3 describes how the sojourn time distribution is calculated for a given customer in an $M/M/s$ queue with $k \geq 0$ customers initially present in the system. Section 4 includes examples using the implemented procedures to calculate exact sojourn time distributions, related measures of performance, and graphical illustrations for varying parameters such as traffic intensity and number of customers in the system. Section 5 offers two approaches for calculating the covariance and correlation among customers in an $M/M/1$ queue. Sections 6 and 7 extend the covariance and correlation calculations by automating the process of finding the joint probability density function of two sojourn times, and provide the exact covariance and correlation calculations for varying traffic intensities and account for occasions where customers are present at time 0. Section 8 concludes the paper with a review of the content and offers areas of further study. Commented code is available for all computations conducted here.

2. Basics of the $M/M/s$ Queue

The $M/M/s$ queue is governed by independent and identically distributed (iid) exponential interarrival times (the arrival stream is a Poisson process) with arrival rate λ and iid exponential service times among s identical servers, each with service rate μ. The interarrival times and the service times are mutually independent. The traffic intensity of the system is $\rho = \lambda/s\mu$. The system consists of a single queue with customers waiting to be serviced by one of the s identical parallel servers. If an arriving customer finds at least one idle server, the customer immediately proceeds to service; otherwise, the customer joins the single queue of those waiting for service in a first-come, first-served manner. To achieve classic steady-state results, the traffic intensity must satisfy $\rho < 1$. This critical assumption is not required in transient analysis described here, because the system of interest never reaches equilibrium.

Let $P_k(n, i)$ be the probability that, upon the arrival of the nth customer, there are i customers in the system including the nth customer (in queue or in service) given k customers are present at time $t = 0$. Using propositions provided by Kelton and Law (1985), reprinted here for completeness (proofs are available in the reference) as well as a recursion algorithm, $P_k(n, i)$ for $i = 1, 2, \ldots, n+k$ can be computed. Using these probabilities, it is possible to find the distribution of the sojourn time for the nth customer in an $M/M/s$ queue given k customers are present at time $t = 0$. Proposition 1 addresses the case of no exits prior to the nth customer's arrival given $k \geq 1$. Proposition 2 is identical to Proposition 1 except that the system is empty and idle at $t = 0$ (i.e., $k = 0$). Proposition 3 addresses the case that the first customer finds $i - 1$ other customers present for $k > 0$. Proposition 4 is the more general case that customer $n \geq 2$ finds i other customers present given $k \geq 0$.

Proposition 1. *If $k \geq 1$, then for $n \geq 1$,*

$$P_k(n, k+n) = \begin{cases} [\rho/(\rho+1)]^n & \text{if } k \geq s, \\[2ex] \rho^n \Big/ \prod_{j=1}^{n} [\rho + (k+j-1)/s] & \text{if } k+n \leq s, \\[2ex] \rho^n \Big/ \left[(\rho+1)^{n-s+k} \prod_{j=1}^{s-k} [\rho + (k+j-1)/s] \right] & \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } k < s < k+n. \end{cases}$$

Proposition 2. *For $n \geq 1$,*

$$P_0(n, n) = \begin{cases} \rho^n \Big/ \prod_{j=1}^{n} [\rho + (j-1)/s] & \text{if } n \leq s, \\[2ex] \rho^n \Big/ \left[(\rho+1)^{n-s} \prod_{j=1}^{s} [\rho + (j-1)/s] \right] & \text{if } n > s. \end{cases}$$

Proposition 3. *If $k \ge 1$, then for $2 \le i \le k$,*

$$P_k(1,i) = \begin{cases} \{\rho/[\rho+(i-1)/s]\} \\ \quad \cdot \displaystyle\prod_{j=1}^{k-i+1} \{1-\rho/[\rho+(k-j+1)/s]\} & \textit{if } k \le s, \\ \rho/(\rho+1)^{k-i+2} & \textit{if } k > s \textit{ and } i > s, \\ \{\rho/[(\rho+1)^{k-s+1}[\rho+(i-1)/s]]\} \\ \quad \cdot \displaystyle\prod_{j=1}^{s-i} \{1-\rho/[\rho+(s-j)/s]\} & \textit{if } i \le s < k. \end{cases}$$

Proposition 4. *For $n \ge 2$, and $2 \le i \le k+n-1$,*

$$P_k(n,i) = \begin{cases} [\rho/(\rho+1)] \displaystyle\sum_{j=i-1}^{k+n-1} [1/(\rho+1)]^{j-i+1} P_k(n-1,j) \\ \hfill \textit{if } i > s, \\ \{\rho/[\rho+(i-1)/s]\} \\ \quad \cdot \left\{ \displaystyle\sum_{j=i-1}^{s-1} \left[\prod_{h=1}^{j-i+1} \{1-\rho/[\rho+(j-h+1)/s]\} \right] \right. \\ \quad \cdot P_k(n-1,j) + \left[\displaystyle\prod_{h=1}^{s-i} \{1-\rho/[\rho+(s-h)/s]\} \right] \\ \quad \left. \cdot \displaystyle\sum_{j=s}^{k+n-1} [1/(\rho+1)]^{j-s+1} P_k(n-1,j) \right\} & \textit{if } i \le s. \end{cases}$$

Using these four propositions, $P_k(n, 1)$ is calculated by subtracting the complementary probability from one. These results can be generated by invoking the Maple procedure `MMsQprob(n,k,s)`, where

- n is the index of the customer of interest,
- k is the number of customers in the system at time $t = 0$, and
- s is the number of identical parallel servers.

The procedure is written in Maple and uses A Probability Programming Language (APPL), which can be downloaded for free at http://www.APPLsoftware.com and is described in Glen et al. (2001). We choose to calculate the distribution of the sojourn time because it is a purely continuous random variable that enables us to exploit associated procedures in APPL.

3. Creating the Sojourn Time Distribution

Once the necessary $P_k(n, i)$, $i = 1, 2, \ldots, n+k$, probabilities are calculated, the exact sojourn time distribution for the nth customer can be calculated. We define X_n as the number of customers, including customer n, in the system upon the arrival of the nth customer. The possible values of X_n can vary from a minimum of 1, which occurs when customer n arrives to an empty queue with idle servers, to a maximum of $n+k$, which occurs when zero exits occur prior to customer n's arrival, matching the possible values for i in the expression for $P_k(n, i)$. The mathematical derivations for both the $M/M/1$ and $M/M/s$ queues make extensive use of the memoryless property, which permits the construction of the distribution of T_n (the sojourn time of customer n). We present each case separately.

3.1. Distribution of T_n for the $M/M/1$ Queue

For an $M/M/1$ queue starting empty and idle, the delay time of the first customer is zero because the customer proceeds directly to service upon arrival. Therefore, the first customer has an exponential(μ) sojourn time distribution. Conditioning on customer 1's service time, one can calculate the probabilities of customer 2 arriving before and after customer 1 finishes service. These well-known results (Kleinrock 1975, Hillier and Lieberman 2005, Winston 2004) are

$$\Pr(Y < X) = \frac{\mu}{\lambda+\mu}, \quad \Pr(X < Y) = \frac{\lambda}{\lambda+\mu},$$

where X is an exponential(λ) interarrival time and Y is an exponential(μ) service time. The first probability represents customer 2 proceeding directly to service, in which case his sojourn time is simply his service time (exponential(μ)). The second probability represents the likelihood that customer 2 will delay prior to service. Using the memoryless property, customer 2 delays an exponential(μ) time before being serviced in an additional exponential(μ) time. Using these two probabilities, it is easy to see that customer 2's sojourn time distribution is a mixture where the mix probabilities are $P_0(n, i)$ and the distributions are determined by the orderings of delays and services potentially encountered. It is well known that for $X_1, X_2, \ldots, X_n$ iid exponential(λ) random variables

$$\sum_{i=1}^{n} X_i \sim \text{Erlang}(\lambda, n). \tag{1}$$

Using this result, the $M/M/1$ queue sojourn time distribution for $k = 0$ initial customers generalizes very elegantly to include $k > 0$, as indicated in Table 1. Line i of Table 1, where $i = 1, 2, \ldots, n+k$, occurs with probability $P_k(n, i)$ and lists the distribution of the sojourn time for the nth customer, conditioned on i customers being in the system upon his arrival.

Let $g_i(t)$ be the probability density function (PDF) of an Erlang(μ, i) random variable. Using the conditional sojourn time distributions for $i = 1, 2, \ldots, n+k$ potential customers in the system, each with probability $P_k(n, i)$, the PDF for the nth customer's sojourn time T_n is the mixture

$$f_n(t) = \sum_{i=1}^{n+k} P_k(n, i) g_i(t), \quad t > 0. \tag{2}$$

Table 1 **Conditional Sojourn Time Distributions for the $M/M/1$ Queue**

X_n	Delay	Service	Conditional sojourn time distribution
1	0	exponential(μ)	exponential(μ)
2	exponential(μ)	exponential(μ)	Erlang(μ, 2)
3	Erlang(μ, 2)	exponential(μ)	Erlang(μ, 3)
4	Erlang(μ, 3)	exponential(μ)	Erlang(μ, 4)
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n+k$	Erlang(μ, $n+k-1$)	exponential(μ)	Erlang(μ, $n+k$)

This result is simple in the $M/M/1$ case because we can take advantage of (1), which results in a mixture of $n+k$ Erlang distributions.

3.2. Distribution of T_n for the $M/M/s$ Queue

Given $s > 1$ parallel identical servers, the nth customer's sojourn time distribution is still a mixture of $n+k$ conditional sojourn time distributions. However, each distribution might be more complicated. For illustration, consider an $M/M/3$ queue starting empty and idle with exponential(λ) arrivals and three identical exponential(μ) servers. It is clear that for customers 1, 2, and 3, the sojourn time is exponential(μ) because all three customers proceed directly to service. Therefore, in the general case, for the number of customers in the system including customer n (which we defined as X_n), the conditional sojourn time distribution is exponential(μ) when $X_n \le s$. However, if $X_n > s$, then the nth customer experiences a delay while observing $X_n - s$ service completions. When $s > 1$ and $X_n > s$, the service time distribution observed by customers in the queue is exponential with a rate of $s\mu$. Using this result, it is apparent that the delay time for the nth customer is the sum of $X_n - s$ independent exponential($s\mu$) random variables, and using (1) is Erlang($s\mu$, $X_n - s$). To calculate the nth customer's sojourn time for a particular value of X_n, we sum his delay and service times. Table 2 shows the distributions conditioned on the number of customers X_n encountered by customer n (including himself) for the $M/M/3$ queue, given $k = 0$ customers present at time $t = 0$. The APPL procedure `Convolution` calculates the distribution of a sum of independent random variables. We use the symbol $\oplus$ to denote convolution.

Table 2 **Conditional Sojourn Time Distributions for the $M/M/3$ Queue with $k = 0$**

X_n	Delay	Service	Conditional sojourn time distribution
1	0	exponential(μ)	exponential(μ)
2	0	exponential(μ)	exponential(μ)
3	0	exponential(μ)	exponential(μ)
4	exponential(3μ)	exponential(μ)	exponential(3μ) $\oplus$ exponential(μ)
5	Erlang(3μ, 2)	exponential(μ)	Erlang(3μ, 2) $\oplus$ exponential(μ)
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	Erlang(3μ, $n-3$)	exponential(μ)	Erlang(3μ, $n-3$) $\oplus$ exponential(μ)

Since X_n represents the number of customers in the system upon arrival of the nth customer (including himself), the first row in Table 2 corresponds to customer n arriving to an empty system, and the last row corresponds to no service completions prior to customer n's arrival. The general form for the $M/M/s$ sojourn time PDF is identical to (2); however, in the $M/M/s$ case, each $g_i(t)$ can potentially require an additional step to calculate the distribution of a sum of random variables.

4. Transient Analysis Applications

It is apparent that calculating (2) for large n is tedious. Kelton and Law (1985) acknowledge the computational difficulty in achieving the $P_k(n, i)$ probabilities alone. Conducting the added steps of up to $n - s$ convolutions for the $M/M/s$ queue and then mixing the resulting conditional distributions with the appropriate probabilities can be complicated to implement. APPL provides the underlying computational engine to achieve exact results for such problems. The APPL procedure `Queue(X,Y,n,k,s)` returns the exact sojourn time distribution for customer n. `Queue` subsequently calls `MMsQprob(n,k,s)`, which uses recursion to calculate the necessary $P_k(n, i)$ probabilities. APPL is capable of symbolic results, as illustrated in Examples 1 and 2.

Example 1. Consider an $M/M/1$ queue with an arrival rate λ and a service rate μ starting empty and idle at time $t = 0$. For the fourth customer, calculate the probabilities $P_0(4, i)$ for $i = 1, 2, 3, 4$.

The APPL command `MMsQprob(4,0,1)` returns the exact symbolic probabilities

$$P_0(4,1) = \frac{5\rho^2+4\rho+1}{(\rho+1)^5},$$

$$P_0(4,2) = \frac{\rho(5\rho^2+4\rho+1)}{(\rho+1)^5},$$

$$P_0(4,3) = \frac{\rho^2(3\rho+1)}{(\rho+1)^4},$$

$$P_0(4,4) = \frac{\rho^3}{(\rho+1)^3},$$

where $\rho = \lambda/\mu$. It is easy to verify that for any $\rho > 0$, $\sum_{i=1}^{4} P_0(4, i) = 1$, as is required. For example, a simple substitution, letting $\rho = 9/10$, yields

$$P_0(4,1) = \tfrac{865000}{2476099}, \qquad P_0(4,2) = \tfrac{778500}{2476099},$$

$$P_0(4,3) = \tfrac{29970}{130321}, \qquad P_0(4,4) = \tfrac{729}{6859}.$$

Example 2. For the queue described in Example 1, calculate the fourth customer's sojourn time distribution, mean sojourn time, and sojourn time variance.

The APPL statements

```
X := ExponentialRV(lambda);
Y := ExponentialRV(mu);
T := Queue(X,Y,4,0,1);
Mean(T);
Variance(T);
```

calculate the desired results. The first two lines define the interarrival and service time distributions, and the third line calculates the fourth customer's sojourn time distribution. The last two lines are self-explanatory. The resulting PDF is

$$\begin{aligned} f_4(t) = {} & \frac{1}{6(\lambda+\mu)^5}\mu^4 e^{-\mu t} \\ & \cdot (30\lambda^2 + 30\lambda^3 t + 24\lambda\mu + 24\lambda^2\mu t + 6\mu^2 + 6\mu^2\lambda t \\ & \quad + 9t^2\lambda^4 + 12t^2\lambda^3\mu + 3t^2\lambda^2\mu^2 + t^3\lambda^5 \\ & \quad + 2t^3\lambda^4\mu + t^3\lambda^3\mu^2), \quad t > 0. \end{aligned}$$

Using $f_4(t)$, the `Mean` and `Variance` commands return

$$E[T_4] = \frac{\mu^5 + 6\lambda\mu^4 + 26\mu^2\lambda^3 + 16\mu^3\lambda^2 + 17\mu\lambda^4 + 4\lambda^5}{\mu(\lambda+\mu)^5}$$

and

$$\begin{aligned} \mathrm{Var}[T_4] = {} & (181\mu^2\lambda^8 + 484\mu^3\lambda^7 + 816\mu^4\lambda^6 + 868\mu^5\lambda^5 \\ & + 574\mu^6\lambda^4 + 244\mu^7\lambda^3 + 40\mu\lambda^9 + 68\mu^8\lambda^2 \\ & + 12\mu^9\lambda + \mu^{10} + 4\lambda^{10})/(\mu^2(\lambda+\mu)^{10}). \end{aligned}$$

Substituting $\lambda = 1$ and $\mu = 10/9$, for example, the results simplify to

$$\begin{aligned} f_4(t) &= \tfrac{5000}{66854673} e^{-10/9t} \\ & \quad \cdot (361t^3 + 2109t^2 + 5190t + 5190), \quad t > 0, \\ E[T_4] &= \tfrac{23323347}{12380495} \approx 1.8839, \quad \text{and} \\ \mathrm{Var}[T_4] &= \tfrac{383506725720906}{153276656445025} \approx 2.5021. \end{aligned}$$

The CPU time associated with the examples is negligible. Examples 1 and 2 represent simple applications of these procedures that circumvent time intensive hand calculations.

Example 3. Calculate the mean sojourn time of the 30th customer in an $M/M/2$ queue with an arrival rate $\lambda = 1$, a service rate $\mu = 9/20$ ($\rho = 10/9$), and $k = 3$ customers initially present.

The mean can be calculated in a single APPL statement by embedding the function calls:

```
Mean(Queue(ExponentialRV(1),ExponentialRV(9/20),
30,3,2));
```

which yields

$$\begin{aligned} & (2074703020765530930928383247885331056322365205 \\ & \quad 26343624731399405569875101728767946601484880 \\ & \quad 138641283564474794935548876340) \\ & \cdot (21534046672820071947860003352210296689224691 \\ & \quad 678842510431455073374994143953948660661783359 \\ & \quad 707587864512638771645692063053)^{-1}, \end{aligned}$$

or, to four digits, 9.6345.

The ability to represent the sojourn time distribution for the nth customer in closed form also provides valuable information on asymptotic behavior for queueing systems, including steady-state convergence rates for different initial conditions. Figure 1 shows the mean sojourn time for customer $n = 1, 2, \ldots, 120$ in an $M/M/1$ queue with $\lambda = 1$, $\mu = 10/9$, and $\rho = 9/10$ for several values of k. The points that are plotted have been connected by lines. As expected, despite the initial condition, all cases appear to move toward the steady-state value of 9 with increasing n. The horizontal axis is only limited to $n = 120$ for display purposes, and in fact, identical computations were carried out for $n > 300$ customers to verify convergence. However, as shown in the cases where $k = 6$ and $k = 10$, the convergence to steady state is not always monotone. Additionally, in testing various traffic intensities, the rate of convergence to steady state increases rapidly with decreasing traffic intensity for varying values of k.

APPL also has the ability to calculate the closed-form cumulative distribution function (CDF) for the nth customer's sojourn time, which permits CDF comparisons for varying n as well as distribution percentiles for a given customer. The procedure call `CDF(T)` returns the exact CDF for customer 4 (from Example 1). Figure 2 displays the sojourn time CDF for varying n with fixed $k = 0$ and $\rho = 9/10$. The differences in CDFs across n correspond to the increasing mean attributed to the delays experienced by successive customers; e.g., customer 2 has delay time 0 or exponential(μ), whereas the nth customer (for $n > 2$) faces a finite mixture of n potential delay distributions. The CDF associated with $n = \infty$ corresponds

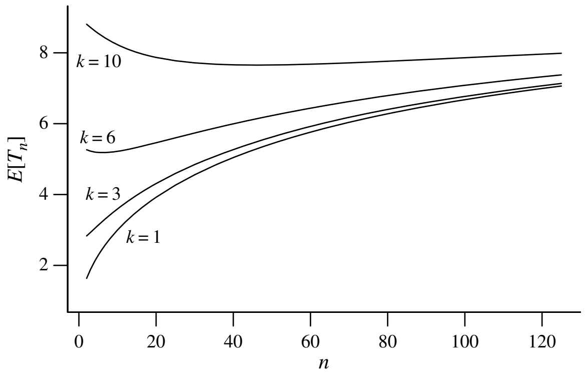

Figure 1 $M/M/1$ **Mean Sojourn Time for** $\rho = 9/10$ **Given** k **at** $t = 0$

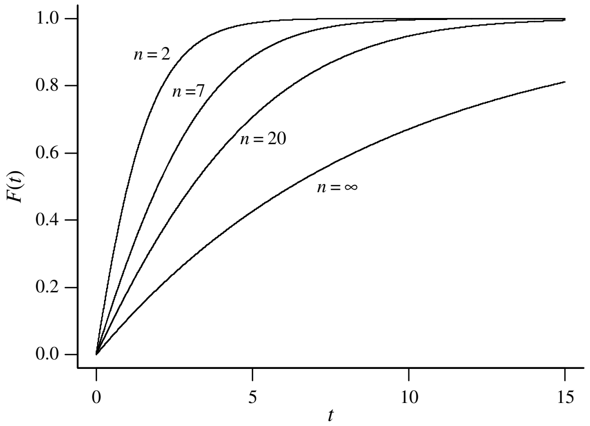

Figure 2 $M/M/1$ **Sojourn Time CDFs for Various** n **Given** $\lambda = 1$, $\mu = 10/9$, $\rho = 9/10$, **and** $k = 0$

to the steady-state distribution of the sojourn time, which is exponentially distributed with a mean of 9 (Kleinrock 1975).

Varying k for an $M/M/1$ queue also provides another basis for comparison of CDFs. Figure 3 fixes $n = 2$ and $\rho = 9/10$, and it plots the resulting CDFs across k. Kelton and Law (1985) make a similar comparison using convergence to steady-state delay time. Using the CDF for multiple values of k allows direct comparison of sojourn time percentiles for customer n. As depicted, the sojourn time CDF for customer 2 is extremely sensitive to the initial condition k. As an illustration, the 80th percentiles for $k = 0, 3$, and 6 are

$$F_2^{-1}(0.80) \approx \begin{cases} 1.935 & k = 0, \\ 4.432 & k = 3, \\ 7.510 & k = 6. \end{cases}$$

These percentiles are achieved using the APPL statements

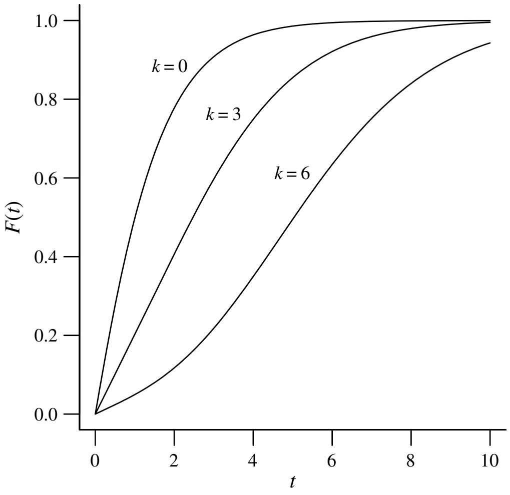

Figure 3 $M/M/1$ **Sojourn Time CDFs for Customer** $n = 2$ **for Various** k **Given** $\lambda = 1$, $\mu = 10/9$, **and** $\rho = 9/10$

```
X := ExponentialRV(1);
Y := ExponentialRV(10/9);
Z := Queue(X,Y,2,k,1);
IDF(Z,0.8);
```

when $k = 0, 3$, and 6. The last statement, `IDF(Z,0.8)`, numerically solves $F_Z(z) = 0.80$ on the interval $(0, \infty)$.

Given the complete specification of the sojourn time distribution, one can use APPL to calculate not only the mean but also the second, third, and fourth moments for customer n. This is especially valuable for steady-state analysis. It is common in simulation to verify attainment of steady-state behavior by examining the mean delay or mean sojourn time. Although some literature exists on estimating transient mean and variance, we are not aware of any literature addressing higher moments. Literature addressing the second moment seems mostly focused on variance estimation and not necessarily convergence. Therefore, even when the first moment might acceptably approximate the steady-state value, there is reason for further analysis of higher moments. For example, Figure 4 displays the first four moments of the sojourn time for customer n in an $M/M/1$ queue, where $\lambda = 1$, $\mu = 2$, and $\rho = 1/2$, with the initial condition $k = 0, 4, 8$.

The code used to calculate the values plotted in Figure 4 is

```
X := ExponentialRV(1);
Y := ExponentialRV(2);
for i from 2 to 60 by 1 do
  T := Queue(X,Y,i,k,1):
  print(i,evalf(Mean(T)), evalf(Variance(T)),
    evalf(Skewness(T)), evalf(Kurtosis(T))):
od:
```

for $k = 0, 4$, and 8. The vertical dashed lines give the smallest customer number for which all three transient values are within 1% of the steady-state value. The relatively low traffic intensity $\rho = 1/2$ was selected purposely to allow quick convergence and easy visual inspection. Even with this somewhat low traffic intensity, it is apparent that the higher moments converge more slowly than the lower moments. In other scenarios where $\rho > 1/2$, the higher moments exhibit an even slower convergence. Each vertical dashed line in Figure 4 was triggered by the $k = 8$ curve, which suggests that the moments are more sensitive to a heavily preloaded system. For the cases $k = 0, 4$, and 8, the customer numbers for which the transient results were within 1% of the steady-state values are listed in Table 3. To verify the initial condition effect on the convergence rate of the first four moments, k was increasingly incremented beyond 8 and displayed a further slowing of convergence.

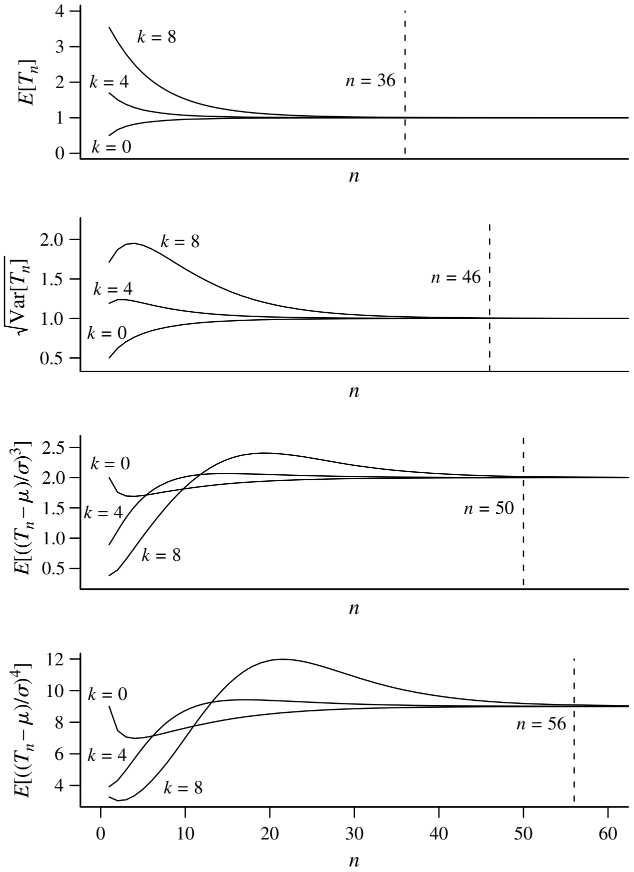

Figure 4 First Four Moments of the $M/M/1$ Sojourn Time for Customers 2–60 for $\rho = 1/2$ and $k = 0, 4, 8$ (the Arrival Rate Is $\lambda = 1$ and the Service Rate Is $\mu = 2$, Resulting in Steady-State Values for the Four Measures of Performance of 1, 1, 2 and 9)

5. Covariance and Correlation in the $M/M/1$ Queue

The dependence exhibited in sojourn times of successive customers is one reason for the difficulty in calculating interval estimators for queue measures of performance. In the simplest case, consider an initially empty and idle $M/M/1$ queue with interarrival and service rates λ and μ, respectively. Our desire is to calculate the covariance between the sojourn times of customers 1 and 2. We outline two approaches to modeling the events in the queue that will be helpful in the presentation of the analytic result.

Table 3 Smallest Customer Number Where the Sojourn Time Transient Result Is Within 1% of Steady State for an $M/M/1$ Queue with $k = 0, 4, 8$ and $\rho = 1/2$

	$k = 0$	$k = 4$	$k = 8$
$E[T]$	19	21	36
$\sqrt{\mathrm{Var}[T]}$	27	29	46
$E[((T-\mu)/\sigma)^3]$	28	29	50
$E[((T-\mu)/\sigma)^4]$	34	35	56

5.1. Discrete-Event Simulation Approaches

As previously discussed, customer 1 proceeds directly to service, and two cases exist for customer 2. In the first case, customer 2 proceeds directly to service. In the second case, he delays proceeding to service until customer 1's departure. Both cases are shown in Figure 5. This section introduces two approaches for generating the first two customers' sojourn times.

The first approach is a standard discrete-event simulation model. Without loss of generality, assume that customer 1 arrives at time 0. In the next-event approach, customer 1's service time is distributed according to the exponential(μ) service time distribution. The arrival time a_2 for customer 2 is distributed according to the exponential(λ) time-between-arrivals distribution. If the arrival occurs after customer 1's service completion (case 1), then customer 2 also has an independent exponential(μ) service time, which in this case is equal to his sojourn time T_2. In the second case where customer 2's arrival time occurs before customer 1's completion of service ($a_2 < T_1$), customer 2 delays for $T_1 - a_2$ time units. We then add the exponential(μ) service time to the delay time to calculate T_2.

We define the gap that occurs in case 2 as $Y = T_1 - a_2$. It can be shown analytically that $Y \sim$ exponential(μ) by computing the distribution of the difference $T_1 - A_2$, where A_2 is the random arrival time of the second customer and is distributed exponential(λ), and then truncating the result on the left at zero. (Alternatively, it can be reasoned that $Y \sim$ exponential(μ) by the memoryless property for the exponential distribution because the remaining service time for customer 1 after customer 2's arrival has the same distribution as an unconditional service time.) Therefore, by using (1), the sojourn time for customer 2 in the second case is distributed Erlang($\mu, 2$).

The second approach is a conditional discrete-event simulation model, where the initial event, whose occurrence time is denoted as E_1 in Figure 6, is either a completion of service for customer 1 with probability $\mu/(\lambda + \mu)$ or the arrival of customer 2 with

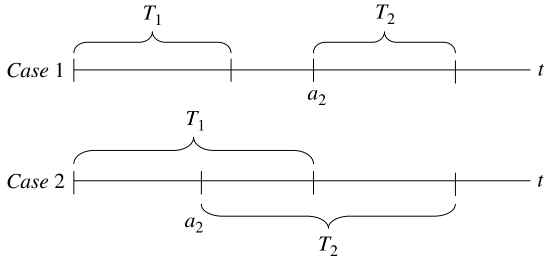

Figure 5 Standard Discrete-Event Simulation Approach for Cases 1 and 2

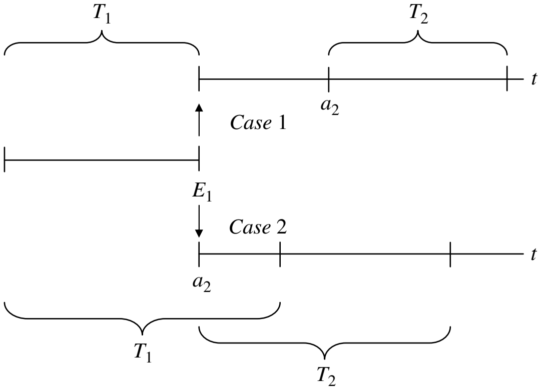

Figure 6 Conditional Discrete-Event Simulation Approach for Cases 1 and 2

probability $\lambda/(\lambda+\mu)$. Since E_1 is the minimum of the arrival time of customer 2 and the service time of customer 1, $E_1 \sim \text{exponential}(\lambda+\mu)$.

5.2. Covariance Calculations

One way to calculate the exact covariance between customers 1 and 2 requires the joint probability density function $f_{T_1,T_2}(t_1,t_2)$. The method used here for computing the joint density applies Theorem 1 below.

Theorem 1. *Let* $X_1 \sim \text{exponential}(\lambda_1)$, $X_2 \sim \text{exponential}(\lambda_2)$, *and* $X_3 \sim \text{exponential}(\lambda_3)$ *be independent random variables. The joint probability density function of* $(T_1, T_2) = (X_1+X_2, X_1+X_3)$ *is*

$$f_{T_1,T_2}(t_1,t_2) = \begin{cases} \dfrac{\lambda_1\lambda_2\lambda_3(e^{\lambda_1 t_1} - e^{(\lambda_2+\lambda_3)t_1})e^{-\lambda_1 t_1-\lambda_2 t_1-\lambda_3 t_2}}{\lambda_1-\lambda_2-\lambda_3} \\ \quad 0 < t_1 < t_2, \\ \dfrac{\lambda_1\lambda_2\lambda_3(e^{\lambda_1 t_2} - e^{(\lambda_2+\lambda_3)t_2})e^{-\lambda_2 t_1-\lambda_1 t_2-\lambda_3 t_2}}{\lambda_1-\lambda_2-\lambda_3} \\ \quad 0 < t_2 < t_1. \end{cases}$$

Proof. The joint CDF of T_1 and T_2 is

$$\begin{aligned} F_{T_1,T_2}(t_1,t_2) &= \Pr(T_1 \le t_1, T_2 \le t_2) \\ &= \Pr(X_1+X_2 \le t_1, X_1+X_3 \le t_2) \\ &= \Pr(X_2 \le t_1 - X_1, X_3 \le t_2 - X_1) \\ &= \int_0^{\min\{t_1,t_2\}} \Pr(X_2 \le t_1 - x_1, \\ &\qquad X_3 \le t_2 - x_1 \mid X_1 = x_1) \\ &\qquad \cdot f_{X_1}(x_1)\,dx_1 \\ &= \int_0^{\min\{t_1,t_2\}} \Pr(X_2 \le t_1 - x_1 \mid X_1 = x_1) \\ &\qquad \cdot \Pr(X_3 \le t_2 - x_1 \mid X_1 = x_1) f_{X_1}(x_1)\,dx_1 \\ &= \int_0^{\min\{t_1,t_2\}} (1-e^{-\lambda_2(t_1-x_1)})(1-e^{-\lambda_3(t_2-x_1)}) \\ &\qquad \cdot \lambda_1 e^{-\lambda_1 x_1}\,dx_1 \\ &= \begin{cases} \displaystyle\int_0^{t_1} (1-e^{-\lambda_2(t_1-x_1)})(1-e^{-\lambda_3(t_2-x_1)}) \\ \quad \cdot \lambda_1 e^{-\lambda_1 x_1}\,dx_1 \quad 0 < t_1 < t_2, \\ \displaystyle\int_0^{t_2} (1-e^{-\lambda_2(t_1-x_1)})(1-e^{-\lambda_3(t_2-x_1)}) \\ \quad \cdot \lambda_1 e^{-\lambda_1 x_1}\,dx_1 \quad 0 < t_2 < t_1. \end{cases} \end{aligned}$$

After evaluating the integrals and differentiating, $f_{T_1,T_2}(t_1,t_2)$ is

$$f_{T_1,T_2}(t_1,t_2) = \begin{cases} \dfrac{\lambda_1\lambda_2\lambda_3(e^{\lambda_1 t_1} - e^{(\lambda_2+\lambda_3)t_1})e^{-\lambda_1 t_1-\lambda_2 t_1-\lambda_3 t_2}}{\lambda_1-\lambda_2-\lambda_3} \\ \quad 0 < t_1 < t_2, \\ \dfrac{\lambda_1\lambda_2\lambda_3(e^{\lambda_1 t_2} - e^{(\lambda_2+\lambda_3)t_2})e^{-\lambda_2 t_1-\lambda_1 t_2-\lambda_3 t_2}}{\lambda_1-\lambda_2-\lambda_3} \\ \quad 0 < t_2 < t_1. \quad \square \end{cases}$$

Theorem 1 provides the joint PDF of the first two sojourn times for case 2, which must be weighted appropriately by the probability that the arrival of customer 2 occurs prior to customer 1's completion of service, or $\lambda/(\lambda+\mu)$. Using the conditional discrete-event model, case 1 consists of independent sojourn times; thus the joint density can be written as the product of the densities of the sojourn times $T_1 \sim \text{exponential}(\lambda+\mu)$ and $T_2 \sim \text{exponential}(\mu)$ and weighted by $\mu/(\lambda+\mu)$. The resulting joint density is a mixture of the two possible cases displayed in Figure 6. We apply Theorem 1 to case 2 because of the dependence that occurs as a result of the overlap of the sojourn times. Figure 7 depicts the relationships between the sojourn times T_1 and T_2 and the random variables X_1, X_2, and X_3 used in Theorem 1.

Substituting $\lambda_1 = \mu$, $\lambda_2 = \lambda+\mu$, and $\lambda_3 = \mu$ into the mixture of cases 1 and 2 yields the joint PDF of T_1 and T_2 as

$$f_{T_1,T_2}(t_1,t_2) = \begin{cases} \dfrac{\mu^2(\lambda e^{-\mu t_2} + \mu e^{-\lambda t_1-\mu t_1-\mu t_2})}{\lambda+\mu} \\ \quad 0 < t_1 < t_2, \\ \dfrac{\mu^2(\lambda e^{-\lambda t_1-\mu t_1+\lambda t_2} + \mu e^{-\lambda t_1-\mu t_1-\mu t_2})}{\lambda+\mu} \\ \quad 0 < t_2 < t_1. \end{cases} \tag{3}$$

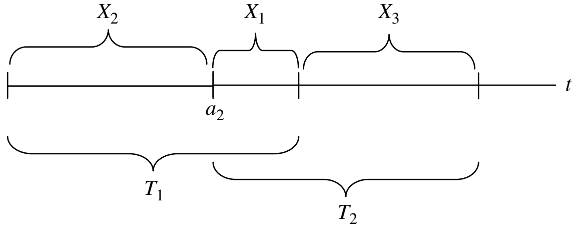

Figure 7 Case 2 for Theorem 1 with $X_1 \sim \text{exponential}(\lambda_1)$, $X_2 \sim \text{exponential}(\lambda_2)$, and $X_3 \sim \text{exponential}(\lambda_3)$

Using this joint PDF, the covariance between the sojourn times of customers 1 and 2 is

$$\mathrm{Cov}(T_1, T_2) = \frac{\lambda(\lambda+2\mu)}{(\lambda+\mu)^2\mu^2}.$$

Substituting $\lambda = 1$ and $\mu = 2$, for example, produces

$$\mathrm{Cov}(T_1, T_2) = \tfrac{5}{36} \approx 0.1389.$$

We now use the results of Theorem 1 in Example 4.

Example 4. Let T_1 and T_2 be the sojourn times for customers 1 and 2, respectively, in an initially empty and idle $M/M/1$ queue with exponential($\lambda = 1$) times between arrivals and exponential($\mu = 2$) service times. Find the distribution of the sample mean $\overline{T} = (T_1 + T_2)/2$ as well as $E[\overline{T}]$ and $\mathrm{Var}[\overline{T}]$.

Applying Equation (3) with $\lambda = 1$ and $\mu = 2$, the joint PDF of T_1 and T_2 is

$$f_{T_1,T_2}(t_1, t_2) = \begin{cases} \frac{8}{3}e^{-3t_1-2t_2} + \frac{4}{3}e^{-2t_2} & 0 < t_1 < t_2, \\ \frac{8}{3}e^{-3t_1-2t_2} + \frac{4}{3}e^{-3t_1+t_2} & 0 < t_2 < t_1. \end{cases}$$

Define the transformation

$$U = \overline{T} = (T_1 + T_2)/2 \quad \text{and} \quad V = (T_1 - T_2)/2$$

with inverse

$$T_1 = U + V \quad \text{and} \quad T_2 = U - V.$$

It can be shown that the functions U and V define a one-to-one transformation; thus, using the bivariate transformation technique described in Hogg et al. (2005), the joint PDF of U and V is

$$f_{U,V}(u, v) = \begin{cases} f_{T_1,T_2}(u+v, u-v)|J| & -u \le v < 0, \\ f_{T_1,T_2}(u+v, u-v)|J| & 0 < v < u, \end{cases}$$

where J is the Jacobian of the inverse transformation defined as

$$J = \begin{vmatrix} \dfrac{\partial t_1}{\partial u} & \dfrac{\partial t_1}{dv} \\ \dfrac{\partial t_2}{\partial u} & \dfrac{\partial t_2}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = -2.$$

Substituting $t_1 = u + v$, $t_2 = u - v$, $J = -2$, and integrating out the dummy transformation variable v, the resulting PDF of $U = \overline{T}$ is

$$f_U(u) = 4e^{-4u} + 2e^{-2u} - 6e^{-6u}, \quad u > 0.$$

The mean of U is

$$\begin{aligned} E[U] &= \int_0^\infty u \cdot f_U(u) du \\ &= \int_0^\infty u \cdot (4e^{-4u} + 2e^{-2u} - 6e^{-6u})\, du \\ &= \tfrac{7}{12}. \end{aligned}$$

Likewise, the variance of U using $\mathrm{Var}[U] = E[U^2] - (E[U])^2$, where

$$\begin{aligned} E[U^2] &= \int_0^\infty u^2 \cdot f_U(u)\, du \\ &= \int_0^\infty u^2 \cdot (4e^{-4u} + 2e^{-2u} - 6e^{-6u})\, du \\ &= \tfrac{41}{72}, \end{aligned}$$

results in

$$\mathrm{Var}[U] = \tfrac{41}{72} - \left[\tfrac{7}{12}\right]^2 = \tfrac{11}{48}.$$

Using the `Queue(X,Y,n,k,s)` procedure for customers 1 and 2, the mean sojourn times are $E[T_1] = 1/2$ and $E[T_2] = 2/3$, and the corresponding variances are $\mathrm{Var}[T_1] = 1/4$ and $\mathrm{Var}[T_2] = 7/18$, respectively. The covariance of sojourn times T_1 and T_2 was identified as $\mathrm{Cov}(T_1, T_2) = 5/36$. Therefore, the mean sojourn time for customers 1 and 2 is

$$E\left[\frac{T_1 + T_2}{2}\right] = \frac{E[T_1] + E[T_2]}{2} = \frac{7}{12},$$

and the variance is

$$\mathrm{Var}\left[\frac{T_1 + T_2}{2}\right] = \frac{\mathrm{Var}[T_1] + \mathrm{Var}[T_2] + 2\mathrm{Cov}(T_1, T_2)}{4} = \frac{11}{48},$$

confirming the moments of $U = \overline{T}$ given above.

Proceeding in this manner, we now derive similar expressions for the first three customers arriving to an initially empty and idle $M/M/1$ queue. We could use first principles to derive the trivariate PDF $f_{T_1,T_2,T_3}(t_1, t_2, t_3)$; however, because covariance only occurs between two customers, it is easier to calculate each respective paired joint distribution for covariance calculations. When considering $n = 3$ customers, there are five possible arrival and departure orderings. In general, for n customers, the number of ways arrivals and departures can occur is given by the nth Catalan number (Stanley 1999), which is

$$C_n = \frac{(2n)!}{n!(n+1)!}.$$

Figure 8 shows the five possible arrangements for $n = 3$ customers along with the sojourn times T_1, T_2, and T_3 for each. The arrival and completion times for the ith customer are denoted by a_i and c_i, respectively. The vertical arrows at event times represent service completions (pointing up) or arrivals (pointing down). This competing-event approach parallels the second discrete-event simulation approach from §5.1. Using the same conditioning approach as in the proof of Theorem 1, the joint PDFs for each of the pairs (T_1, T_2), (T_1, T_3), and (T_2, T_3) in each of the five cases can be determined and then mixed to achieve the three associated joint PDFs. The mixture probabilities

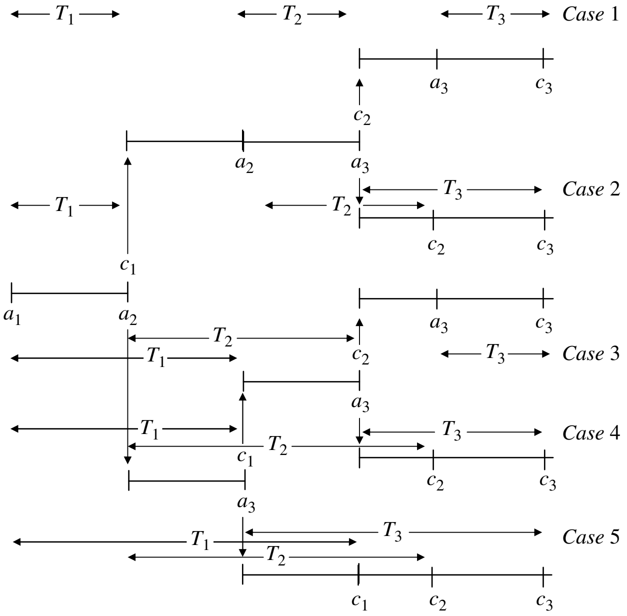

Figure 8 Five Cases for $n = 3$ Customers' Sojourn Times in an $M/M/1$ Queue

are calculated by multiplying the appropriate number of competing arrivals (with probability $\lambda/(\lambda+\mu)$) or service completions (with probability $\mu/(\lambda+\mu)$). For example, in case 1 shown in Figure 8, there are two instances with competing risks, both of which result in a service completion; thus the probability of this case is $\mu^2/(\lambda+\mu)^2$. Using these joint densities, the symmetric variance–covariance matrix for the first $n = 3$ customer sojourn times

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

is depicted in Table 4.

Substituting $\lambda = 1$ and $\mu = 2$, for example, results in

$$\Sigma = \begin{bmatrix} 1/4 & 5/36 & 29/324 \\ \bullet & 7/18 & 13/54 \\ \bullet & \bullet & 1451/2916 \end{bmatrix}$$

$$\approx \begin{bmatrix} 0.2500 & 0.1389 & 0.0895 \\ \bullet & 0.3889 & 0.2407 \\ \bullet & \bullet & 0.4976 \end{bmatrix}.$$

The sojourn time variance increases with customer number down the diagonal of the matrix because of the nature of the queueing process, where the sojourn time distribution for each additional customer is dependent on all the previous customers. On the other hand, the off-diagonal covariance entries in each row decrease with customer separation; for example, $\sigma_{13} < \sigma_{12}$.

Table 4 Sojurn Time Variance–Covariance Matrix for the First $n = 3$ Customers in an $M/M/1$ Queue

$\dfrac{1}{\mu^2}$	$\dfrac{\lambda(2\mu+\lambda)}{(\lambda+\mu)^2\mu^2}$	$\dfrac{\lambda^2(\lambda^2+4\lambda\mu+5\mu^2)}{(\lambda+\mu)^4\mu^2}$
$\bullet$	$\dfrac{2\lambda^2+4\lambda\mu+\mu^2}{(\lambda+\mu)^2\mu^2}$	$\dfrac{\lambda(2\lambda^2+8\lambda^2\mu+11\lambda\mu^2+2\mu^3)}{(\lambda+\mu)^4\mu^2}$
$\bullet$	$\bullet$	$\dfrac{3\lambda^6+18\lambda^5\mu+45\lambda^4\mu^2+54\lambda^3\mu^3+30\lambda^2\mu^4+8\lambda\mu^5+\mu^6}{(\lambda+\mu)^6\mu^2}$

6. Extending Covariance Calculations

Consider the $n = 3$ case, where all three customers arrive prior to the first customer's completion of service (this is case 5 in Figure 8). Using a "1" to represent an arrival and a "−1" for a departure, this sequence of arrivals and departures can be represented by the vector

$$\begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}.$$

Figure 9 depicts this case as a path from left to right, where moving up and right indicates an arrival and moving down and right indicates a service completion. Horizontal moves are not permitted. Each of the five possible sequences of arrivals and departures for $n = 3$, shown in Figure 8, can be depicted by a specific path from the bottom left node to the bottom right node. The number of customers in the system is depicted by the height of each node in a path in Figure 9.

Ruskey and Williams (2008) present an elegant algorithm that generates all such paths of arrival and service completions for a given number of customers n. The algorithm is based on a simple iterative successor rule that uses prefix shifts to exhaust the possible arrival and service completion scenarios. In Figure 9 these are the $6!/(3!4!) = 5$ paths that can be drawn from the bottom left node to the bottom right node. The algorithm is "loopless" in that it requires a constant amount of computation in transforming the current case to its successor. Define the case matrix C

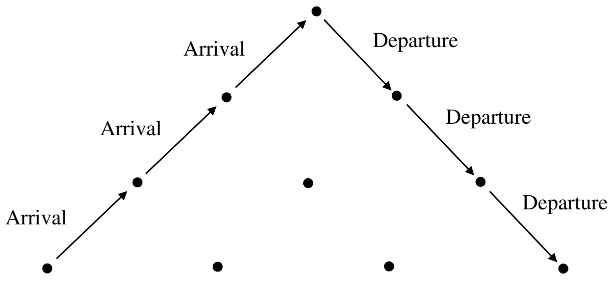

Figure 9 Path for Case 5 of $n = 3$ Customers Arrival and Departure Pattern in an $M/M/1$ Queue

with dimension $(2n)!/(n!(n+1)!)$ by $2n$ as the exhaustive list of possible arrival and service completion scenarios for n customers. To initiate the matrix, the first row of C is

$$C_1 = \begin{bmatrix} 1 & -1 & 1 & 1 & -1 & -1 \end{bmatrix}.$$

The first row is always the ordered string created by an arrival, a service completion, $n-1$ arrivals, and $n-1$ service completions. The iterative successor rule described by Ruskey and Williams (2008, p. 107) is "Locate the leftmost $[-1,\ 1]$ and suppose its 1 is in position k. If the $(k+1)$-st prefix shift is valid (a possible arrival/service completion sequence), then it is the successor; if it is not valid then the k-th prefix shift is the successor." The $(k+1)$st prefix shift for the sequence

$$B = \{B_1, B_2, \ldots, B_{k-1}, B_k, B_{k+1}, \ldots, B_{2n}\}$$

is defined to be

$$B = \{B_1, B_{k+1}, B_2, \ldots, B_{k-1}, B_k, B_{k+2}, \ldots, B_{2n}\};$$

that is, the $(k+1)$st element of the sequence is shifted into the second position, and the relative order of the other elements is left unchanged. The length of the sequence is always $2n$ because the number of arrivals and departures is balanced at n each. An example of an invalid sequence is

$$\begin{bmatrix} 1 & -1 & -1 & 1 & 1 & -1 \end{bmatrix}$$

because the second service completion occurs prior to the second arrival. For $n = 3$, the case matrix C is

$$C = \begin{bmatrix} 1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}.$$

(Note that the order of the five rows does not match the order of the cases in Figure 8.)

Figure 10 further categorizes each segment of the path based on whether there exists a competing risk (competing event), in which case the distribution of the time until the next event (either an arrival or a completion) is given by

$$\min\{\text{exponential}(\lambda), \text{exponential}(\mu)\}$$
$$\sim \text{exponential}(\lambda+\mu),$$

where the time between arrivals is distributed as exponential(λ) and the service time distribution is exponential(μ).

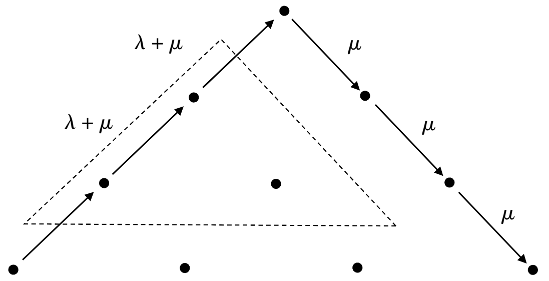

Figure 10 **Path Segment Distributions for Case 5 for $n = 3$ Customers**

Competing risks can only occur along path segments that originate inside the dashed triangle shown in Figure 10. These path segments are exponential($\lambda+\mu$) distributed and are correspondingly labeled $\lambda+\mu$. Once all customers have arrived, the only possible events are service completions; thus each path segment along the rightmost edge of Figure 10 is distributed exponential(μ) and labeled μ. If the path of interest reaches a node at the bottom of the figure, the queueing system empties, and the next event must be an arrival, which occurs in an exponential(λ) time into the future. While the system is empty, none of the customers' sojourn times are affected; therefore waiting for the next arrival does not affect customer sojourn time distribution. The interior triangle in the path diagram also provides a method to calculate the probability of all possible paths. For path segments originating inside the triangle, a move right and up occurs with probability $\lambda/(\lambda+\mu)$, and a move right and down occurs with probability $\mu/(\lambda+\mu)$. For the particular path shown in Figure 10, there are two segments originating inside the triangle, both of which are right and up, thus representing two successive arrivals. Therefore the probability of this case is

$$\frac{\lambda}{\lambda+\mu} \cdot \frac{\lambda}{\lambda+\mu} = \frac{\lambda^2}{(\lambda+\mu)^2}.$$

To capture the structure of the segment distributions for a given path, represented as a row of the case matrix C, another vector of length $2n-1$ is created where each entry corresponds to the sojourn time distribution for a particular segment. There are three possible entries in this vector:

1. exponential($\lambda+\mu$), which is indicated by a 1;
2. exponential(μ), which is indicated by a 2; and
3. no distribution as a result of an emptied system, which is depicted as a 0.

The vector is of length $2n-1$ because the first customer's arrival time can be ignored; it does not affect sojourn time. For the particular path shown in Figure 10, the corresponding segment distribution vector is

$$\begin{bmatrix} 1 & 1 & 2 & 2 & 2 \end{bmatrix}.$$

Define the new matrix C' with dimension $(2n)!/(n!(n+1)!)$ by $2n-1$ as the segment distribution matrix for each case in C. For $n=3$, the matrix C' is

$$C' = \begin{bmatrix} 1 & 0 & 1 & 2 & 2 \\ 1 & 1 & 1 & 2 & 2 \\ 1 & 0 & 1 & 0 & 2 \\ 1 & 1 & 1 & 0 & 2 \\ 1 & 1 & 2 & 2 & 2 \end{bmatrix}.$$

The two vectors (which are each the fifth row of the corresponding matrices),

$$C_5 = \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix} \quad \text{and}$$
$$C'_5 = \begin{bmatrix} 1 & 1 & 2 & 2 & 2 \end{bmatrix},$$

contain the information necessary to calculate the contribution of case 5 to the joint PDF for the sojourn times of any two customers. Denote C_l as the lth row vector of case matrix C, and define the 2×2 matrix R_l with elements

$$R_l = \begin{bmatrix} r_{is} & r_{if} \\ r_{js} & r_{jf} \end{bmatrix},$$

where r_{is} and r_{if} are the start and finish indices for customer i in row l of the case matrix C, respectively. Define r_{js} and r_{jf} similarly for customer j. Using C_5 above, for customers $i=1$ and $j=3$,

$$R_5 = \begin{bmatrix} 1 & 4 \\ 3 & 6 \end{bmatrix}.$$

Customer 1′s arrival is the first event to occur. Customer 1′s departure is the fourth event to occur. Customer 3′s arrival is the third event to occur. Customer 3′s departure is the sixth event to occur.

The R_l matrix provides two critical pieces of information. First, for the given case l, if $r_{if} < r_{js}$, then the sojourn times for customers i and j do not overlap because customer i departs prior to customer j's arrival. Because in each specific case the sojourn time for each customer comprises a uniquely determined sequence of independent time segments, consisting of either service completions distributed exponential(μ) or interarrival times distributed exponential($\lambda+\mu$), and because the sequences for customers i and j have no time segments in common, the sojourn times for customers i and j are independent. Therefore, if $r_{if} < r_{js}$, the contribution of case l to the joint PDF is created by simply multiplying the sojourn time PDFs for customers i and j. Second, by computing $r_{if} - r_{is}$ and $r_{jf} - r_{js}$ and then indexing across C'_l, the appropriate segment distributions can be combined to form the joint sojourn time PDF for customers i and j.

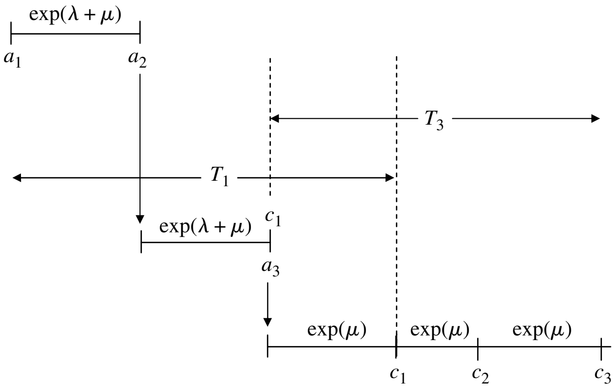

Figure 11 Sojourn Time Segments for Customers 1 and 3 in Case 5 for $n=3$ Customers

When $r_{if} > r_{js}$, the joint probability distribution is calculated by conditioning in a similar fashion to the proof of Theorem 1. However, it is first necessary to find the independent and overlapping segments for the customers of interest. For the arrival and service completion scenario described by C_5, Figure 11 shows sojourn times T_1 and T_3 for customers 1 and 3. The independent portion of customer 1's sojourn time consists of the two exponential($\lambda+\mu$) segments. The independent portion of customer 3's sojourn time consists of the two exponential(μ) segments, shown on the right side of Figure 11. The dependent (overlap) portion between customers 1 and 3 consists of the single exponential(μ) segment falling within the dashed vertical lines. Using C'_5 and R_5, these segments can be determined without reference to Figure 11 as follows: given $r_{1f} > r_{3s}$, (that is, customer 3 arrives prior to customer 1 completing service) the independent portions of customer 1's sojourn time distribution are found by (a) calculating $r_{3s} - r_{1s} = 3 - 1 = 2$ and then (b) collecting the elements in C'_5, beginning at index $r_{1s} = 1$ and indexing $r_{3s} - r_{1s} - 1 = 1$ additional element of the vector. For $C'_5 = \begin{bmatrix} 1 & 1 & 2 & 2 & 2 \end{bmatrix}$, the first two entries, c'_{51} and c'_{52}, correspond to the two exponential($\lambda+\mu$) segments. Likewise, customer 3's independent sojourn time segments are found by (a) calculating $r_{3f} - r_{1f} = 6 - 4 = 2$ and then (b) collecting the elements in C'_5, beginning at index $r_{1f} = 4$ and indexing $r_{3f} - r_{1f} - 1 = 1$ additional element of the vector. This amounts to the two exponential(μ) segments in elements 4 and 5 of C'_5. The dependent portion is identified by starting at the element $r_{3s} = 3$ and indexing $r_{if} - r_{3s} - 1 = 0$ additional elements, which is the third element of C'_5, a single exponential(μ) segment. In this case, calculating the joint PDF is straightforward because the independent portions for each customer are iid exponential random variables. Defining the independent cumulative distribution function portions for customers 1 and 3 as $X_1 \sim \text{Erlang}(\lambda+\mu, 2)$ and $X_3 \sim \text{Erlang}(\mu, 2)$, respectively, and the dependent (overlap) random variable as $W \sim$ exponential(μ), the contribution of case 5 to the joint CDF of $(T_1, T_3) = (X_1 + W, X_3 + W)$, conditioning on the dependent distribution segment W, is

$$
\begin{aligned}
F_{T_1,T_3}(t_1,t_3) &= \Pr(T_1 \le t_1, T_3 \le t_3)\\
&= \Pr(X_1 + W \le t_1, X_3 + W \le t_3)\\
&= \Pr(X_1 \le t_1 - W, X_3 \le t_3 - W)\\
&= \int_0^{\min\{t_1,t_3\}} \Pr(X_1 \le t_1 - w,\\
&\qquad\qquad X_3 \le t_3 - w \mid W = w)\\
&\qquad\qquad \cdot f_W(w)\,dw\\
&= \int_0^{\min\{t_1,t_3\}} \Pr(X_1 \le t_1 - w \mid W = w)\\
&\quad \cdot \Pr(X_3 \le t_3 - w \mid W = w) f_W(w)\,dw\\
&= \int_0^{\min\{t_1,t_3\}} F_{X_1}(t_1 - w) F_{X_3}(t_3 - w)\mu e^{-\mu w}\,dw\\
&= \begin{cases}
\displaystyle\int_0^{t_1} F_{X_1}(t_1 - w) F_{X_3}(t_3 - w)\mu e^{-\mu w}\,dw\\
\qquad 0 < t_1 < t_3,\\
\displaystyle\int_0^{t_3} F_{X_1}(t_1 - w) F_{X_3}(t_3 - w)\mu e^{-\mu w}\,dw\\
\qquad 0 < t_3 < t_1.
\end{cases}
\end{aligned}
$$

Because closed-form versions of $F_{X_1}(t_1 - w)$ and $F_{X_3}(t_3 - w)$ are available, Maple is capable of evaluating this expression; for large n, however, it can be time consuming.

When the independent distribution segments are not iid exponential random variables, the calculation is more problematic because we can no longer use (1) to easily express $F_{X_1}(t_1 - w)$ and $F_{X_3}(t_3 - w)$. Convolution is required, and though capable, Maple, and subsequently APPL, slow very quickly with increasing n. To overcome this shortfall, let us consider Theorem 2, which appears to be a faster approach than the two suggested in Hagwood (2009).

Theorem 2. *If $S_1 \sim$Erlang(λ_1, m) and $S_2 \sim$Erlang(λ_2, n) are independent random variables, then the PDF of $Y = S_1 + S_2$ is*

$$
\begin{aligned}
f_Y(y) = \Bigg[\frac{\lambda_1^m \lambda_2^n e^{-\lambda_2 y}}{(m-1)!(n-1)!} \sum_{x=0}^{n-1} \Bigg\{ (-1)^x \binom{n-1}{x} y^{n-1-x} e^{(\lambda_2 - \lambda_1)s}\\
\cdot \sum_{r=0}^{m-1+x} (-1)^r \frac{(m-1+x)!\, s^{m-1+x-r}}{(m-1+x-r)!(\lambda_2 - \lambda_1)^{r+1}} \Bigg\}\Bigg]_{s=0}^{y},\\
y > 0.
\end{aligned}
$$

Proof. Since S_1 and S_2 are independent, the PDF of $Y = S_1 + S_2$ using convolution and the binomial theorem is

$$
\begin{aligned}
f_Y(y) &= \int_0^y f_{S_1}(s) f_{S_2}(y-s)\,ds\\
&= \int_0^y \frac{\lambda_1(\lambda_1 s)^{m-1} e^{-\lambda_1 s}}{(m-1)!} \frac{\lambda_2(\lambda_2(y-s))^{n-1} e^{-\lambda_2(y-s)}}{(n-1)!}\,ds\\
&= \frac{\lambda_1^m \lambda_2^n}{(m-1)!(n-1)!} \int_0^y s^{m-1} e^{-\lambda_1 s} (y-s)^{n-1} e^{-\lambda_2 (y-s)}\,ds\\
&= \frac{\lambda_1^m \lambda_2^n e^{-\lambda_2 y}}{(m-1)!(n-1)!} \int_0^y s^{m-1} (y-s)^{n-1} e^{s(\lambda_2 - \lambda_1)}\,ds\\
&= \frac{\lambda_1^m \lambda_2^n e^{-\lambda_2 y}}{(m-1)!(n-1)!}\\
&\quad \cdot \int_0^y s^{m-1} \left(\sum_{x=0}^{n-1} \binom{n-1}{x} y^{n-1-x} (-s)^x \right) e^{s(\lambda_2 - \lambda_1)}\,ds\\
&= \frac{\lambda_1^m \lambda_2^n e^{-\lambda_2 y}}{(m-1)!(n-1)!} \sum_{x=0}^{n-1} \Bigg\{ (-1)^x \binom{n-1}{x} y^{n-1-x}\\
&\qquad\qquad \cdot \int_0^y s^{m-1+x} e^{s(\lambda_2 - \lambda_1)}\,ds \Bigg\}\\
&= \Bigg[\frac{\lambda_1^m \lambda_2^n e^{-\lambda_2 y}}{(m-1)!(n-1)!} \sum_{x=0}^{n-1} \Bigg\{ (-1)^x \binom{n-1}{x} y^{n-1-x} e^{(\lambda_2 - \lambda_1)s}\\
&\quad \cdot \sum_{r=0}^{m-1+x} (-1)^r \frac{(m-1+x)!\, s^{m-1+x-r}}{(m-1+x-r)!(\lambda_2 - \lambda_1)^{r+1}} \Bigg\}\Bigg]_{s=0}^{y},\\
&\qquad\qquad y > 0. \quad \square
\end{aligned}
$$

The APPL procedure `Cov(a,b,n)` applies Theorem 2 to calculate the covariance between the sojourn times of customers a and b $(a < b)$ in a system of n customers. For computational considerations (i.e., evaluating the fewest cases necessary for a given n), setting the number of customers $n = b$ provides the fastest result. Additionally, calling `Cov(a,b,n)` where $n > b$ produces a result identical to $n = b$ because customers arriving after customer b do not affect the covariance of previous customers.

Rewriting the integral as a sum via Theorem 2 avoids the calls to `Convolution(X,Y)` in APPL as well as the need to integrate for each case and piece. One can always use this approach, even when the independent part of a particular customer's sojourn time contains many independent distribution segments. The times for these segments can only be exponential($\lambda + \mu$) distributed or exponential(μ) distributed, which implies that their sum can always be written as the sum of two independent Erlang random variables. This approach speeds computation time considerably. The symmetric variance–covariance matrix for $n = 10$ customers with parameters $\lambda = 1$, $\mu = 2$, and $\rho = 1/2$ is showcased in Table 5; exact values are provided.

CPU time is a factor in these computations. Each element in the 10th column of the variance–

Table 5 **Sojourn Time Variance–Covariance Matrix for the First $n = 10$ Customers in an $M/M/1$ Queue with $\lambda = 1$, $\mu = 2$ (Exact Values)**

$\frac{1}{4}$	$\frac{5}{36}$	$\frac{29}{324}$	$\frac{181}{2916}$	$\frac{1181}{26244}$	$\frac{2647}{78732}$	$\frac{18191}{708588}$	$\frac{127111}{6377292}$	$\frac{2699837}{172186884}$	$\frac{19319845}{1549681956}$
•	$\frac{7}{18}$	$\frac{13}{54}$	$\frac{239}{1458}$	$\frac{1543}{13122}$	$\frac{10303}{118098}$	$\frac{23485}{354294}$	$\frac{163493}{3188646}$	$\frac{3462503}{86093442}$	$\frac{24719519}{774840978}$
•	•	$\frac{1451}{2916}$	$\frac{8531}{26244}$	$\frac{53995}{236196}$	$\frac{356291}{2125764}$	$\frac{805705}{6377292}$	$\frac{5576849}{57395628}$	$\frac{39197977}{516560652}$	$\frac{836647331}{13947137604}$
•	•	•	$\frac{34514}{59049}$	$\frac{209794}{531441}$	$\frac{1357010}{4782969}$	$\frac{3031606}{14348907}$	$\frac{20810726}{129140163}$	$\frac{145390102}{1162261467}$	$\frac{3088887890}{31381059609}$
•	•	•	•	$\frac{12525605}{19131876}$	$\frac{77889229}{172186884}$	$\frac{170586983}{516560652}$	$\frac{1156711327}{4649045868}$	$\frac{8013045911}{41841412812}$	$\frac{169183999981}{1129718145924}$
•	•	•	•	•	$\frac{551583889}{774840978}$	$\frac{1162296371}{2324522934}$	$\frac{7727099083}{20920706406}$	$\frac{52871149859}{188286357654}$	$\frac{1106749378225}{5083731656658}$
•	•	•	•	•	•	$\frac{10582107143}{13947137604}$	$\frac{67728246079}{125524238436}$	$\frac{454382575415}{1129718145924}$	$\frac{9394007745229}{30502389939948}$
•	•	•	•	•	•	•	$\frac{225196533287}{282429536481}$	$\frac{1455144635743}{2541865828329}$	$\frac{29498588275973}{68630377364883}$
•	•	•	•	•	•	•	•	$\frac{75890492486993}{91507169819844}$	$\frac{1482244865480580}{2470693585135780}$
•	•	•	•	•	•	•	•	•	$\frac{28549065408995300}{33354363399333100}$

Table 6 **Sojourn Time Variance–Covariance Matrix for the First $n = 10$ Customers in an $M/M/1$ Queue with $\lambda = 1$, $\mu = 2$ (Approximations)**

0.2500	0.1389	0.0895	0.0621	0.0450	0.0336	0.0257	0.0199	0.0157	0.0125
•	0.3889	0.2407	0.1639	0.1176	0.0872	0.0663	0.0513	0.0402	0.0319
•	•	0.4976	0.3251	0.2286	0.1676	0.1263	0.0972	0.0759	0.0600
•	•	•	0.5845	0.3948	0.2837	0.2113	0.1611	0.1251	0.0984
•	•	•	•	0.6547	0.4524	0.3302	0.2488	0.1915	0.1498
•	•	•	•	•	0.7119	0.5000	0.3694	0.2808	0.2177
•	•	•	•	•	•	0.7587	0.5396	0.4022	0.3080
•	•	•	•	•	•	•	0.7974	0.5725	0.4298
•	•	•	•	•	•	•	•	0.8293	0.5999
•	•	•	•	•	•	•	•	•	0.8559

Table 7 **Sojourn Time Variance–Covariance Matrix for the First $n = 10$ Customers in an $M/M/1$ Queue with $\lambda = 1$, $\mu = 10/9$ (Approximations)**

0.8100	0.5856	0.4737	0.4040	0.3553	0.3189	0.2904	0.2673	0.2481	0.2318
•	1.3956	1.1097	0.9393	0.8226	0.7363	0.6692	0.6150	0.5702	0.5323
•	•	1.9561	1.6298	1.4167	1.2626	1.1441	1.0494	0.9714	0.9057
•	•	•	2.5021	2.1458	1.8995	1.7142	1.5679	1.4484	1.3485
•	•	•	•	3.0364	2.6565	2.3831	2.1715	2.0009	1.8593
•	•	•	•	•	3.5605	3.1614	2.8652	2.6310	2.4389
•	•	•	•	•	•	4.0754	3.6600	3.3444	3.0904
•	•	•	•	•	•	•	4.5818	4.1524	3.8199
•	•	•	•	•	•	•	•	5.0803	4.6386
•	•	•	•	•	•	•	•	•	5.5713

covariance matrix is calculated from a joint PDF that is a mixture of $C_{10} = 20!/(10!11!) = 16{,}796$ component distributions, each corresponding to a unique ordering of arrivals and departures.

Because these values are difficult to compare in fractional form, the same matrix is provided again, with matrix elements rounded to four decimal places; see Table 6.

As the traffic intensity increases, so do the values in the variance–covariance matrix. To illustrate, the same matrix is provided for the increased traffic intensity parameters $\lambda = 1$, $\mu = 10/9$, and $\rho = 9/10$. The increasing sojourn time variance along the diagonal is expected with the increasing traffic intensity. In addition, the rate that covariance between customers decreases as customer separation increases is less pronounced; see Table 7.

Using this variance–covariance matrix for traffic intensity $\rho = 9/10$, let us consider the following example.

Example 5. Let T_i, $i = 1, 2, \ldots, 10$, be the sojourn times for the first $n = 10$ customers in an $M/M/1$

Table 8 Sojourn Time Variance–Covariance Matrix for the First $n = 10$ Customers in an $M/M/1$ Queue with $\lambda = 1$, $\mu = 2/3$ (Approximations)

2.2500	1.8900	1.7172	1.6135	1.5438	1.4937	1.4558	1.4263	1.4027	1.3835
•	4.1400	3.7368	3.5018	3.3459	3.2344	3.1507	3.0856	3.0337	2.9913
•	•	6.0957	5.6825	5.4166	5.2292	5.0896	4.9817	4.8958	4.8261
•	•	•	8.1312	7.7208	7.4397	7.2332	7.0747	6.9493	6.8479
•	•	•	•	10.2424	9.8410	9.5538	9.3361	9.1652	9.0276
•	•	•	•	•	12.4235	12.0342	11.7463	11.5230	11.3444
•	•	•	•	•	•	14.6687	14.2931	14.0081	13.7828
•	•	•	•	•	•	•	16.9727	16.6115	16.3319
•	•	•	•	•	•	•	•	19.3310	18.9846
•	•	•	•	•	•	•	•	•	21.7397

queue, with arrival rate $\lambda = 1$ and service rate $\mu = 10/9$, that is initially empty and idle. Find the variance of the average sojourn time for the first 10 customers.

Define the average sojourn time as

$$\overline{T} = \tfrac{1}{10}\sum_{i=1}^{10} T_i.$$

Because the sojourn times are not independent random variables, the variance of the average sojourn time is

$$\begin{aligned}\mathrm{Var}[\overline{T}] &= \mathrm{Var}\left[\tfrac{1}{10}\sum_{i=1}^{10} T_i\right]\\ &= \tfrac{1}{100}\mathrm{Var}\left[\sum_{i=1}^{10} T_i\right]\\ &= \tfrac{1}{100}\left[\sum_{i=1}^{10}\mathrm{Var}[T_i] + 2\sum_{i<j}\sum \mathrm{Cov}(T_i, T_j)\right].\end{aligned}$$

The result is the sum of all elements in the variance–covariance matrix in Table 7 multiplied by the constant 1/100. The sum of the variance–covariance matrix rounded to four significant digits is 177.6642; therefore the variance of $\overline{T}$ is

$$\mathrm{Var}[\overline{T}] \approx 1.7766.$$

To verify the calculation a Monte Carlo simulation was conducted five times, using one million replications each time. The resulting 95% confidence interval for the variance of $\overline{T}$ was $\overline{T} \in (1.773, 1.781)$, which agrees with the analytic result.

Traditional steady-state queueing theory and analysis lacks the insight provided in these transient variance–covariance matrices. For businesses where the number of customers in a day is so small that true steady state is never achieved, routine queueing measures of performance are not representative of reality. Additionally, consider a system where the traffic intensity exceeds 1. For such a system, an analyst might be interested in customer covariance. Increasing the traffic intensity so that $\rho > 1$ does not preclude covariance calculations using this method and therefore allows transient analysis of such systems. A variance–covariance matrix for $\lambda = 1$, $\mu = 2/3$, and $\rho = 3/2$ is presented in Table 8. Given this traffic intensity, the system is unstable, and the expected sojourn times for successive customers increase without bound. Along the main diagonal the customer variance is clearly increasing, and the covariance decreases as the separation occurs between customers. This decrease is monotonic, and although not studied in detail here, it appears that the rate of covariance decrease might be of interest for an unstable traffic intensity.

7. Sojourn Time Covariance with k Customers Initially Present

When k customers are present in the $M/M/1$ queue at time 0, the approach used to compute sojourn time covariance between customers becomes more difficult. When the two customers of interest possess indices larger than k (i.e., T_i where $i > k$), then the approach is similar to that derived in §6. However, there are two other possibilities. The first possibility is that the first customer has an index of k or less, and the second customer has an index larger than k. In this instance, the only difference in deriving the joint CDF is that the lower-indexed customer begins his sojourn time at time 0. In the second possibility, both customers have an index of k or below. If these indices are i and j, where $i < j \le k$, the time intervals for sojourn times T_i and T_j begin at 0. It is obvious that $T_i \le T_j$, because the completion time for customer i must occur prior to the completion time for customer j. For each of these possibilities, the covariance derivation that follows will mirror the empty and idle covariance derivation in §6. To illustrate the calculations, consider an $M/M/1$ queue with $k = 2$ customers initially present at time 0 and a single additional customer, $n = 1$. The transition diagram where the first event (not including the k customers initially present at time 0) is an arrival, which is analogous to Figures 9 and 10, is given in Figure 12. The total number of customers passing through the system is $n + k = 3$. Using a "1" to denote an arrival and a "-1" to denote a departure, each arrival/departure ordering instance for $n + k = 3$ customers must contain exactly three -1s

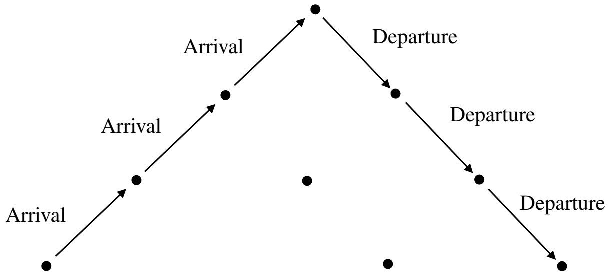

Figure 12 **Transition Diagram for $n+k=1+2=3$ Customers When the First Event Is an Arrival**

(completions of service) and a single 1 (arrival). The algorithm presented by Ruskey and Williams (2008) does not facilitate the listing of all orderings for an unbalanced system, where the number of departures is greater than the number of arrivals (as opposed to an empty and idle queue at time 0). However, we can produce all possible arrival/departure sequences with a simple manipulation of the algorithm as well as count the number of possible sequences. The number of possible orderings, denoted by $C(n \mid k)$, follows, where n represents the number of customers passing through the system that arrive after time 0, and k is the number of customers present at time 0:

$$C(n \mid k) = \sum_{j=0}^{\lfloor k/2 \rfloor} (-1)^j \binom{k-j}{j} C_{n+k-j}$$

for $k = 0, 1, 2, \ldots$ and $n = 1, 2, \ldots,$ where $\lfloor \cdot \rfloor$ denotes the greatest integer function. The case matrix C is found by applying the Ruskey and Williams (2008) algorithm for $n+k$ customers and then deleting the instances where the first k events do not correspond to arrivals. As seen previously, the case matrix for $n+k=1+2=3$ customers is

$$C = \begin{bmatrix} 1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}.$$

Rows 2, 4, and 5 correspond to the first $k=2$ events being arrivals. Rows 1 and 3 must be deleted from the case matrix, because for each row, a completion of service occurs prior to the first two arrivals. Deleting these rows results in the case matrix

$$C = \begin{bmatrix} 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix},$$

with the remaining rows representing all possible arrival/departure sequences. We can further simplify the case matrix by deleting the first k columns, resulting in

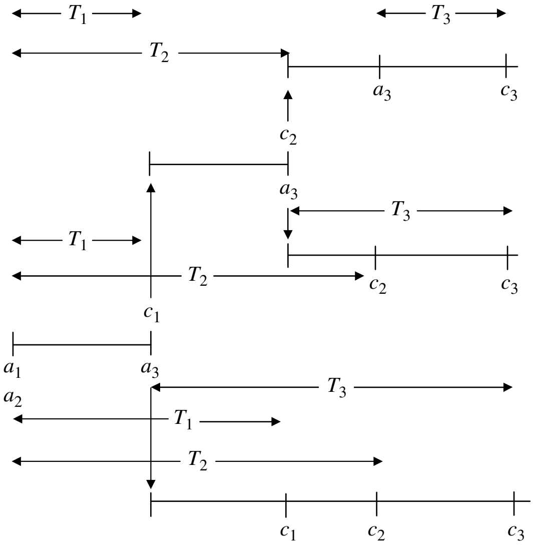

Figure 13 **Three Cases for $k=2$ Initial Customers and a Single $n=1$ Additional Customer in an $M/M/1$ Queue**

$$C = \begin{bmatrix} -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}.$$

The rows of the case matrix correspond to the three cases shown in Figure 13.

The algorithm for computing the joint PDF, and subsequently the covariance, of the sojourn times of any two customers does not differ significantly from the algorithm presented in §6. However, for the sojourn times T_1 and T_2 in Figure 13, a new theorem is introduced.

Theorem 3. *Let $X \sim$ exponential(λ_1) and $Y \sim$ exponential(λ_2) be independent random variables. The joint PDF of $(T_1, T_2) = (X, X+Y)$ is*

$$f_{T_1,T_2}(t_1, t_2) = \lambda_1 \lambda_2 e^{-\lambda_2 t_1 - \lambda_1 t_2 + \lambda_1 t_1} \quad 0 < t_1 < t_2.$$

Proof. The joint CDF of T_1 and T_2 is

$$\begin{aligned} &F_{T_1,T_2}(t_1, t_2) \\ &\quad = \Pr(T_1 \le t_1, T_2 \le t_2) \\ &\quad = \Pr(X \le t_1, X+Y \le t_2) \\ &\quad = \Pr(X \le t_1, Y \le t_2 - X) \\ &\quad = \int_0^{t_1} \int_0^{t_2-x} f_X(x) \cdot f_Y(y)\, dy\, dx \\ &\quad = \int_0^{t_1} \int_0^{t_2-x} (\lambda_1 e^{-\lambda_1 x}) \cdot (\lambda_2 e^{-\lambda_2 y})\, dy\, dx \\ &\quad = \frac{\lambda_1 - \lambda_2 + \lambda_2 e^{-\lambda_1 t_2} + \lambda_2 e^{-\lambda_2 t_1} - \lambda_1 e^{-\lambda_2 t_1} - \lambda_2 e^{-\lambda_2 t_1 - \lambda_1 t_2 + \lambda_1 t_1}}{\lambda_1 - \lambda_2}, \end{aligned}$$

for $0 < t_1 < t_2$. Taking partial derivatives, $f_{T_1,T_2}(t_1, t_2)$ is

$$f_{T_1,T_2}(t_1, t_2) = \lambda_1 \lambda_2 e^{-\lambda_2 t_1 - \lambda_1 t_2 + \lambda_1 t_1} \quad 0 < t_1 < t_2. \qquad \square$$

Theorem 3 provides the joint PDF for the sojourn times T_1 and T_2 of the first two customers initially present at time 0. It may be more complicated to calculate the joint PDFs for the sojourn times of other pairs of customers who were initially present at time 0. This is because if $(i, j) \neq (1, 2)$ and $i < j \leq k$, where k is the number of customers present at time 0, the time intervals of duration X and Y during which customers i and j, respectively, are served may each be composed of multiple independent, exponentially distributed time segments. Each of these multiple segments is limited to only one of two possibilities, an exponential$(\lambda+\mu)$ segment or an exponential(μ) segment. In this more complicated situation, we let $(T_i, T_j) = (X, X+Y)$, as in Theorem 3, and apply Theorem 2 to find the PDFs of X and Y (using the procedure `conv(m,n)`); we then let Maple handle the sojourn time joint PDF calculation. When the second customer of interest has an index greater than or equal to k, the sojourn time joint PDF follows an application of Theorem 1 as described in §6, when cases exist with dependence.

Using the final case matrix C above, the associated segment distribution matrix C' is

$$C' = \begin{bmatrix} 1 & 1 & 2 & 2 \\ 1 & 1 & 0 & 2 \\ 1 & 2 & 2 & 2 \end{bmatrix},$$

where the possible elements are the same as defined in §6. The probability vector associated with the case matrix C is

$$\begin{bmatrix} \frac{2}{9} & \frac{4}{9} & \frac{1}{3} \end{bmatrix}$$

for arrival rate $\lambda = 1$ and service rate $\mu = 2$.

Using the case matrix C and the segment distribution matrix C', the joint PDFs for each case are created by selecting the appropriate segments for a given pair of customers, where the segments are identified by the R_l matrix discussed in §6. Once the joint PDFs are created for each case, they are mixed with the probability vector to determine the sojourn time joint PDF for covariance calculations. These calculations are coded in Maple as the procedure `kCov(X,Y,a,b,n,k)`. The first two arguments X and Y are the distribution of time between arrivals, exponential(λ), and the service time distribution, exponential(μ), respectively. They are entered in the APPL list-of-lists format. The arguments a and b are the customers of interest for the covariance calculation, where $a < b$. The argument n is the number of customers processing through the system not including those present at time 0, which is indicated by the last argument k. Therefore, the total number of customers processing through the system is $n+k$, and a covariance calculation between any two of these customers can be achieved with the appropriate function call. For example, the function call `kCov(ExponentialRV(1), ExponentialRV(2), 1,2, 1,3)` calculates the covariance between customers 1 and 2 in an $M/M/1$ queue with an arrival rate $\lambda = 1$, service time rate $\mu = 2$, three customers present at time 0, and a single additional customer processing through the system. The variance–covariance matrix for an $M/M/1$ queue with an arrival rate $\lambda = 1$ and service rate $\mu = 2$, where $k = 4$ customers are present at time 0 and an additional $n = 6$ customers process through the system, is presented in Table 9.

Unlike the previous variance–covariance matrices, some row elements—in particular, those elements associated with customers who are initially present—do not decrease monotonically. To explain these entries, consider Theorem 4.

Theorem 4. *If* $X_1, X_2, \ldots, X_n$ *are iid* exponential(μ) *random variables and*

$$T_s = \sum_{r=1}^{s} X_r \quad s = 1, 2, \ldots, n,$$

then $\mathrm{Var}(T_i) = \mathrm{Cov}(T_i, T_l)$, $0 < i < l \leq n$.

Proof. Note that $E[T_k] = k/\mu$ for $k = 1, 2, \ldots, n$ and that T_i and X_r are independent for $1 \leq i < r \leq n$:

$$\begin{aligned}
\mathrm{Cov}(T_i, T_l) &= E\left[\left(T_i - \frac{i}{\mu}\right)\left(T_l - \frac{l}{\mu}\right)\right] \\
&= E\left[\left(T_i - \frac{i}{\mu}\right)\left\{\left(T_i - \frac{i}{\mu}\right) + \sum_{r=i+1}^{l}\left(X_r - \frac{1}{\mu}\right)\right\}\right] \\
&= E\left[\left(T_i - \frac{i}{\mu}\right)^2\right] \\
&\quad + E\left\{\sum_{r=i+1}^{l}\left[\left(T_i - \frac{i}{\mu}\right)\left(X_r - \frac{1}{\mu}\right)\right]\right\} \\
&= \mathrm{Var}[T_i] + \sum_{r=i+1}^{l} E\left[T_i - \frac{i}{\mu}\right] E\left[X_r - \frac{1}{\mu}\right] \\
&= \mathrm{Var}[T_i]. \quad \square
\end{aligned}$$

We can apply Theorem 4 to those customer pairs where both indices $i, j \leq k$. Therefore, the entries in the variance–covariance matrix for customer pairs $(1,2)$, $(1,3)$, and $(1,4)$ are

$$\mathrm{Var}[T_1] = \mathrm{Cov}(T_1, T_2) = \mathrm{Cov}(T_1, T_3) = \mathrm{Cov}(T_1, T_4) = \tfrac{1}{4}.$$

Likewise, for the customer pairs $(2,3)$ and $(2,4)$,

$$\mathrm{Var}[T_2] = \mathrm{Cov}(T_2, T_3) = \mathrm{Cov}(T_2, T_4) = \tfrac{1}{2}.$$

Furthermore, it can be shown that, in general,

$$\mathrm{Var}[T_i] = \mathrm{Cov}(T_i, T_j) = \frac{i}{\mu^2}$$

for $i < j \leq k$, where k customers are present at time 0. For example, consider a single-server box office with exponential(μ) service times that will be

Table 9 **Sojourn Time Variance–Covariance Matrix for the First $n = 6$ Customers in an $M/M/1$ Queue with $k = 4$ Customers Initially Present and $\lambda = 1$, $\mu = 2$ (Exact Values)**

$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{211}{972}$	$\frac{1579}{8748}$	$\frac{11651}{78732}$	$\frac{28553}{236196}$	$\frac{630131}{6377292}$	$\frac{4646155}{57395628}$
•	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{211}{486}$	$\frac{1579}{4374}$	$\frac{11651}{39366}$	$\frac{28553}{118098}$	$\frac{630131}{3188646}$	$\frac{4646155}{28697814}$
•	•	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{211}{324}$	$\frac{1579}{2916}$	$\frac{11651}{26244}$	$\frac{28553}{78732}$	$\frac{630131}{2125764}$	$\frac{4646155}{19131876}$
•	•	•	1	$\frac{211}{243}$	$\frac{1579}{2187}$	$\frac{11651}{19683}$	$\frac{28553}{59049}$	$\frac{630131}{1594323}$	$\frac{4646155}{14348907}$
•	•	•	•	$\frac{37289}{26244}$	$\frac{271153}{236196}$	$\frac{1966777}{2125764}$	$\frac{1588153}{2125764}$	$\frac{34755203}{57395628}$	$\frac{763875281}{1549681956}$
•	•	•	•	•	$\frac{1629655}{1062882}$	$\frac{11663887}{9565938}$	$\frac{9353743}{9565938}$	$\frac{203800469}{258280326}$	$\frac{4465399991}{6973568802}$
•	•	•	•	•	•	$\frac{263490131}{172186884}$	$\frac{208262483}{172186884}$	$\frac{4506205633}{4649045868}$	$\frac{98323535707}{125524238436}$
•	•	•	•	•	•	•	$\frac{63939878}{43046721}$	$\frac{1359189250}{1162261467}$	$\frac{29402061622}{31381059609}$
•	•	•	•	•	•	•	•	$\frac{179260456277}{125524238436}$	$\frac{379721786263}{3389154437772}$
•	•	•	•	•	•	•	•	•	$\frac{62708955663745}{45753584909922}$

offering tickets to a popular concert the next day. If 1,000 patrons, each buying one ticket, camp out the night before to get the best seats for the concert, these $k = 1{,}000$ customers are present at time 0, and therefore we can predetermine the covariance between the sojourn times of any two of the customers. Additionally, Theorem 4 presents the counterintuitive result that $\text{Cov}(T_1, T_2) = \text{Cov}(T_1, T_{1,000})$. As expected, the correlation decreases with increasing lag because of the diminishing effect of the intermediate customer sojourn times reflected in the denominator of the defining formula for correlation.

8. Conclusion

Previous transient analysis results for the $M/M/1$ and $M/M/s$ queues have been combined with the functionality of the Maple computational engine (and subsequently APPL) to develop both symbolic and numeric exact sojourn time PDFs that can be manipulated to compute and study various measures of performance. A complete variance–covariance matrix for the first $n = 10$ customers and varying traffic intensity is calculated, which illustrates this approach's ability to determine the joint PDF between two customer sojourn times. The results offer a framework for describing how the well-known $M/M/s$ queue steady-state results occur as the queue progresses toward steady state. When possible, results are checked against corresponding Monte Carlo simulation and/or previous literature. The first principle's derivation suggests that a viable alternative for future research would be to apply the approaches provided in this work to a $G/G/1$ queue. Computational considerations take priority as n increases. Making use of other computational formulae (such as Hagwood 2009) may offer significant time savings and is another interesting avenue for future work.

Acknowledgments

The authors thank Bert Zwart for describing the conditional discrete-event model in Figure 6. They thank Sheldon Ross for offering his assistance in the covariance and correlation sections of this paper. The authors also thank Sid Lawrence, Barry Nelson, Raghu Pasupathy, and Bruce Schmeiser for their assistance in deriving the sojourn time distributions.

References

Abate, J., W. Whitt. 1988. Transient behavior of the $M/M/1$ queue via Laplace transforms. *Adv. Appl. Probab.* **20**(1) 145–178.

de Souza e Silva, E. S., H. R. Gail, R. V. Campos. 1995. Calculating transient distributions of cumulative reward. B. D. Gaither, ed. *Proc. 1995 ACM SIGMETRICS Joint Internat. Conf. Measurement Model. Comput. Systems*, ACM, New York, 231–240.

Gafarian, A. V., C. J. Ancker Jr., T. Morisaku. 1976. The problem of the initial transient in digital computer simulation. *Proc. 76 Bicentennial Conf. Winter Simul., Gaithersburg, MD*, 49–51.

Glen, A. G., D. L. Evans, L. M. Leemis. 2001. APPL: A probability programming language. *Amer. Statistician* **55**(2) 156–166.

Grassmann, W. K. 1977. Transient solutions in Markovian queueing systems. *Comput. Oper. Res.* **4**(1) 47–53.

Grassmann, W. K. 2008. Warm-up periods in simulation can be detrimental. *Probab. Engrg. Inform. Sci.* **22**(3) 415–429.

Hagwood, C. 2009. An application of the residue calculus: The distribution of the sum of nonhomogeneous gamma variates. *Amer. Statistician* **63**(1) 37–39.

Hillier, F. S., G. J. Lieberman. 2005. *Introduction to Operations Research*. McGraw-Hill, New York.

Hogg, R. V., A. T. Craig, J. McKean. 2005. *Introduction to Mathematical Statistics*. Macmillan, New York.

Kelton, W. D. 1985. Transient exponential—Erlang queues and steady-state simulation. *Comm. ACM* **28**(7) 741–749.

Kelton, W. D., A. M. Law. 1985. The transient behavior of the $M/M/s$ queue, with implications for steady-state simulation. *Oper. Res.* **33**(2) 378–396.

Kleinrock, L. 1975. *Queueing Systems*. John Wiley & Sons, New York.

Law, A. M. 1975. A comparison of two techniques for determining the accuracy of simulation output. Technical Report 75–11, University of Wisconsin–Madison, Madison.

Leguesdron, P., J. Pellaumail, G. Rubino, B. Sericola. 1993. Transient analysis of the $M/M/1$ queue. *Adv. Appl. Probab.* **25**(3) 702–713.

Morisaku, T. 1976. Techniques for data-truncation in digital computer simulation. Ph.D. thesis, University of Southern California, Los Angeles.

Odoni, A. R., E. Roth. 1983. An empirical investigation of the transient behavior of stationary queueing systems. *Oper. Res.* **31**(3) 432–455.

Parthasarathy, P. R. 1987. A transient solution to an $M/M/1$ queue: A simple approach. *Adv. Appl. Probab.* **19**(4) 997–998.

Pegden, C. D., M. Rosenshine. 1982. Some new results for the $M/M/1$ queue. *Management Sci.* **28**(7) 821–828.

Ruskey, F., A. Williams. 2008. Generating balanced parentheses and binary trees by prefix shifts. J. Harland, P. Manyem, eds. *Proc. 14th Sympos. Comput.: Australasian Theory*, Vol. 77. Australian Computer Society, Darlinghurst, NSW, Australia, 107–115.

Stanley, R. P. 1999. *Enumerative Combinatorics: Volume 2*. Cambridge Studies in Advanced Mathematics, Vol. 62. Cambridge University Press, Cambridge, UK.

Winston, W. L. 2004. *Operations Research: Applications and Algorithms*. Thomson, Belmont, CA.