Transform-Free Analysis of M/G/1/K and Related Queues

Shun-Chen Niu[1]
School of Management
The University of Texas at Dallas
P. O. Box 830688
Richardson, Texas 75083-0688

Robert B. Cooper[1]
Department of Computer Science and Engineering
Florida Atlantic University
P. O. Box 3091
Boca Raton, Florida 33431-0991

September 1989
Revision: November 1991[2]

[1]Research supported in part by the National Science Foundation under grant DDM-9001751.
[2]*Mathematics of Operations Research*, to appear.

Abstract

Using constructive, sample-path arguments, we derive a variety of transform-free results about queue lengths and waiting times for the $M/G/1/K$ queue. In classical analyses of $M/G/1/K$, it is typical to work with Markov processes obtained by defining the "state" of the system at a time epoch to be the number of customers present and, as supplementary information, the remaining service time of the customer, if any, in service. In contrast, the key idea behind our analysis is to work with a modified Markov process that has a more-detailed state description: At any time epoch t when the server is busy, we replace "the number of customers present" by two variables, namely (a) the number of customers who were (and still are) waiting in the queue immediately after the start of the service in progress, and (b) the number of customers who arrived during that same service but prior to t. We show that this minor change of state definition, coupled with a rigorous formalization of the intuitive notion of a "test customer" (whose viewpoint is adopted in our analysis of the modified Markov process), makes possible a surprisingly simple analysis of the $M/G/1/K$ queue. We also show that our method can be extended easily to yield similar results for several generalizations of the basic $M/G/1/K$ model.

AMS 1980 subject classification. Primary: 90B22; Secondary: 60K25.
IAOR 1973 subject classification. Main: Queues.
OR/MS Index 1978 subject classification. Primary: 681 Queues.

Key words. $M/G/1$ queue, finite capacity, test customer, sample-path analysis, exceptional first services, server vacations, semi-Markovian services.

1 Introduction

Although the $M/G/1/K$ queue has been studied extensively (for $K = \infty$, in particular) for over sixty years, there are still new and interesting insights to be discovered. In this paper, we derive, without using any transforms, a variety of explicit results about queue lengths and waiting times for the $M/G/1/K$ queue. The majority of our results are new as stated, but many can be related to previously-known results in transform form. For most of the related transform results that we have examined, the explicit inversions of the transforms appear to be difficult (although recent work has shown that *numerical* inversions of transforms can be remarkably easy; see, e.g., Abate and Whitt [1992] for a comprehensive review).

In the standard $M/G/1/K$ queue, it is assumed that: The arrival process is Poisson at rate λ; the service times are identically distributed random variables, independent of the arrival process and each other, following distribution function G with mean $1/\mu$; and the system has a total capacity of K customers, including the one, if any, in service. An arriving customer enters the system only if at least one of the $K-1$ waiting positions is available; otherwise, the customer is lost immediately, without receiving any service. When waiting-time distributions are considered, we also assume that entering customers are served in the order of their arrival.

For $t \geq 0$, let $L(t)$ be the number of customers in the system at time t; and, when $L(t) > 0$, let $R(t)$ be the remaining service time of the customer in service. Define

$$\mathbf{Z}(t) = \begin{cases} 0 & \text{if } L(t) = 0\,, \\ \\ (L(t),\, R(t)) & \text{if } L(t) > 0\,; \end{cases} \tag{1.1}$$

then, the Markov process $\mathbf{Z} \equiv \{\mathbf{Z}(t),\, t \geq 0\}$ (or related processes obtained by embedding at arrival and/or departure epochs) is the typical starting point of classical analyses of the $M/G/1/K$ queue (e.g., Cohen [1982], Chapters II.4 and III.6; Keilson [1966]; Kendall [1951, 1953]; Takács [1963]; and Wishart [1961]). In contrast, the key idea behind the analysis in our paper is to "decompose" the variable $L(\cdot)$ in $\mathbf{Z}$ into the sum of three variables (see (1.3) below), whenever it is positive: For any time epoch t at which the server is busy, let $Q_s(t)$ be the number of customers waiting in queue immediately after the start of the current service, and let $Q_a(t)$ be the number of customers who arrived during the current service interval but prior to t, including those, if any, who did *not* enter; then, we shall work with the more-detailed Markov process $\mathbf{Z}^+ \equiv \{\mathbf{Z}^+(t),\, t \geq 0\}$ defined by

$$\mathbf{Z}^+(t) = \begin{cases} 0 & \text{if } L(t) = 0\,, \\ \\ (Q_s(t),\, Q_a(t),\, R(t)) & \text{if } L(t) > 0\,; \end{cases} \tag{1.2}$$

Observe that if $L(t) > 0$, then, since the number of waiting positions equals $K-1$, we have

$$L(t) = 1 + Q_s(t) + \min\left[Q_a(t),\, K-1-Q_s(t)\right]\,, \tag{1.3}$$

where the first term 1 accounts for the customer in service; thus, using (1.3), we can indeed recover (1.1) easily from (1.2).

We will analyze the *continuous-time* process $\mathbf{Z}^+$ from the viewpoint of a "randomly-selected" arriving customer (Niu [1988], pp. 160–162): At a given time epoch t, we say that $\mathbf{Z}^+$ is in state Θ_0^+ if $L(t) = 0$; and that it is in state $\Theta_{ij}^+(x)$ if $L(t) > 0$, $Q_s(t) = i$, $Q_a(t) = j$, and $R(t) \le x$, where $0 \le i \le K-2$, $0 \le j < \infty$, and $x \ge 0$. We also assume, for convenience, that sample paths of $\mathbf{Z}^+$ are left-continuous at arrival epochs, so that, with $\{A_k,\, k \ge 1\}$ denoting customer-arrival epochs, $\mathbf{Z}^+(A_k)$ is the state of $\mathbf{Z}^+$ as seen by the k^{th} arriving customer. Then, we shall study the following long-run averages (or proportions) associated with $\mathbf{Z}^+$:

$$\alpha_0^+ \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\Theta_0^+}(\mathbf{Z}^+(A_k)) \tag{1.4}$$

and, for $0 \le i \le K-2$, $0 \le j < \infty$, and $x \ge 0$,

$$\alpha_{ij}^+(x) \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\Theta_{ij}^+(x)}(\mathbf{Z}^+(A_k))\,, \tag{1.5}$$

where $\mathbf{1}_\theta(\cdot)$ denotes the indicator function of a given state θ (and we assume that these limits converge to constants w.p.1, independent of initial conditions). Observe that, for any *given* sample path, the term $1/n$ in (1.4) and (1.5) can be interpreted as the "probability" of selecting any one of the first n arriving customers, and $\mathbf{1}_\theta(\mathbf{Z}^+(A_k))$, $1 \le k \le n$, as the "conditional probability" for the k^{th} customer, if selected, to find $\mathbf{Z}^+$ in state θ; therefore, by letting $n \to \infty$, (1.4) and (1.5), indeed, give the "state distribution" of $\mathbf{Z}^+$ as seen by a randomly-selected arriving customer.

An important reason for working with (1.4) and (1.5), which formalize the intuitive (and fuzzy) notion of a "test" or "tagged" customer frequently encountered in the queueing literature, is that they are explicit averages over sample paths and, as such, facilitate rigorous constructive proofs. In particular, our main result, given as Theorem 1 in Section 2, is a set of formulas that relate, w.p.1, α_0^+ and $\alpha_{ij}^+(x)$ to several other simpler (or easily computable) sample-path averages associated with the variables $Q_s(\cdot)$, $Q_a(\cdot)$, and $R(\cdot)$; moreover, its proof is essentially *deterministic*, in the same spirit as the classical proofs of $L = \lambda W$ (Stidham [1974]) and PASTA (Wolff [1982]).

Obtaining transform-free results has long been of interest in queueing theory (see, e.g., Neuts [1981], pp. 3 and 27; and Neuts [1982]). In this regard, the contribution of our paper, which can be viewed as a continuation of Niu [1988] and Niu and Cooper [1989, 1991], is

to show that the surprisingly minor shift of attention from the process $\mathbf{Z}$ to the process $\mathbf{Z}^+$, coupled with the concept of a randomly-selected arriving customer, makes possible a constructive, transform-free analysis of the $M/G/1/K$ queue. This approach, in particular, circumvents the difficulties that arise when one adapts the standard arguments for infinite-capacity $M/G/1$ (e.g., the ingenious departure-epoch embedded-Markov-chain approach of Kendall [1951, 1953]) to the analysis of *finite*-capacity $M/G/1/K$ (see Neuts [1981], p. 83, for related comments). We will also show that our method, once developed rigorously and understood, can be extended easily to yield explicit results for a variety of generalizations of the basic $M/G/1/K$ model.

The outline of the rest of our paper is as follows. In Section 2, we state all of our results for the $M/G/1/K$ queue and summarize (some of) their connections to previously-known results for this model. In Section 3, we provide proofs for the assertions of Section 2. In Section 4, we derive similar results for three generalizations of the $M/G/1/K$ queue that allow, respectively, exceptional first service in each busy period, server vacations, and semi-Markovian services. Finally, in Section 5, we comment on future work.

2 M/G/1/K

We begin with some preliminary definitions and results, starting with the description of a *service-start-epoch embedded Markov chain* $\mathbf{Q}$: For $k \geq 1$, let Q_k be the number of customers waiting in the queue immediately *after* the k^{th} service-start epoch, and let N_k be the number of customers who arrive *during* the k^{th} service interval. Then, $\mathbf{Q}$ is defined to be the process $\{Q_k,\, k \geq 1\}$, with Q_k determined recursively by

$$Q_{k+1} = \max\{0,\, Q_k + \min[N_k,\, K - 1 - Q_k] - 1\}\ . \tag{2.1}$$

Observe that N_k is independent of Q_k for every $k \geq 1$; and that the sequence $\{N_k,\, k \geq 1\}$ is i.i.d., with

$$a_j \equiv P\{N_1 = j\} = \int_0^\infty \frac{(\lambda y)^j}{j!} e^{-\lambda y}\, dG(y)\,, \qquad 0 \leq j < \infty\,. \tag{2.2}$$

Hence, the process $\mathbf{Q}$ is indeed a Markov chain. (This service-start-epoch embedded Markov chain has been used previously by Keilson [1966] to study queue lengths in $M/G/1/K$; we will use it in a different way.)

All of our results in this section will be expressed in terms of the stationary probabilities of $\mathbf{Q}$: From (2.1) and (2.2), it is easy to see that $\mathbf{Q}$ is irreducible and aperiodic; and therefore, its stationary probabilities $\sigma_0,\, \sigma_1,\, \cdots,\, \sigma_{K-2}$ are, from standard Markov-chain theory (e.g., Theorem 4.3.3 of Ross [1983]), uniquely determined by the equations

$$\sigma_0 = \sigma_0\,(a_0 + a_1) + \sigma_1 a_0\,, \tag{2.3}$$

$$\sigma_j = \sum_{i=0}^{j+1} \sigma_i \, a_{j+1-i} \,, \qquad 1 \le j < K-2 \,, \tag{2.4}$$

$$\sigma_{K-2} = \sum_{i=0}^{K-2} \sigma_i \sum_{j=K-2}^{\infty} a_{j+1-i} \,, \tag{2.5}$$

and the normalization condition,

$$\sum_{j=0}^{K-2} \sigma_j = 1 \,. \tag{2.6}$$

Notice that (2.3), (2.4), and (2.5) are valid only if $K \ge 3$, which we will assume throughout to avoid notational complications (if $K = 1$, 2, then $\sigma_0 = 1$).

That σ_0, σ_1, $\cdots$, σ_{K-2} arise in our results is a consequence (the detailed connection will be fully described in Section 3.2) of monitoring the status of the variable $Q_s(\cdot)$ (see (1.2)) in $\mathbf{Z}^+$ from the viewpoint of a randomly-selected *service-start epoch*: We say that $\mathbf{Z}^+$ is in state $\Theta_{i\cdot}^+(\infty)$, where $0 \le i \le K-2$, at time t if $L(t) > 0$ and $Q_s(t) = i$; and let $\mathbf{1}_{i\cdot}(k)$, $k \ge 1$, be the indicator function of the event that $\mathbf{Z}^+$ is in state $\Theta_{i\cdot}^+(\infty)$ immediately *after* the k^{th} service-start epoch. Then, again from standard Markov-chain theory (e.g., Ross [1983], p. 135, Problem 4.14), we have, w.p.1,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{i\cdot}(k) = \sigma_i \,. \tag{2.7}$$

Paralleling (2.7), our results will also be expressed in terms of averages derived from monitoring the status of the remaining variables $Q_a(\cdot)$ and $R(\cdot)$ in $\mathbf{Z}^+$, but now from the viewpoint of a randomly-selected *blocked customer* (i.e., one who on arrival finds the server busy): We say that $\mathbf{Z}^+$ is in state $\Theta_{\cdot j}^+(x)$, where $0 \le j < \infty$ and $x \ge 0$, at time t if $L(t) > 0$, $Q_a(t) = j$, and $R(t) \le x$; and let $\mathbf{1}_{\cdot j;\, x}(k)$, $k \ge 1$, be the indicator function of the event that, of the N_k customers (all of which are blocked) that arrive during the k^{th} service interval, there is one, namely the $(j+1)^{th}$, who on arrival finds the process $\mathbf{Z}^+$ in state $\Theta_{\cdot j}^+(x)$ (there can be either 1 or 0 such customer). Then, our results will depend (the detailed connection will be fully described in Section 3.2) on

$$\beta_{\cdot j}^+(x) \equiv \lim_{n\to\infty} \frac{\sum_{k=1}^{n} \mathbf{1}_{\cdot j;\, x}(k)}{\sum_{k=1}^{n} N_k} \,, \tag{2.8}$$

that is, the proportion of blocked customers who, on arrival, find $\mathbf{Z}^+$ in state $\Theta_{\cdot j}^+(x)$. Observe that the sequence $\{\mathbf{1}_{\cdot j;\, x}(k),\, k \ge 1\}$ is i.i.d.; and that, by conditioning on the time of the

$(j+1)^{th}$ arrival after a service-start epoch, we have

$$E[\mathbf{1}_{\cdot j;\, x}(1)] = P\{\mathbf{1}_{\cdot j;\, x}(1) = 1\} = \int_0^\infty [G(y+x) - G(y)] \, \frac{(\lambda y)^j}{j!} e^{-\lambda y} \, \lambda \, dy \, . \tag{2.9}$$

Hence, after dividing by n in both the numerator and the denominator on the right-hand side of (2.8) and applying the SLLN (Strong Law of Large Numbers), we have, w.p.1, $\beta^+_{\cdot j}(x) = E[\mathbf{1}_{\cdot j;\, x}(1)]/E(N_1) \equiv \nu_j(x)$, where, upon substituting (2.9) and $E(N_1) = \lambda/\mu \equiv \rho$ (not necessarily less than 1),

$$\nu_j(x) = \mu \int_0^\infty \frac{(\lambda y)^j}{j!} e^{-\lambda y} \, [G(y+x) - G(y)] \, dy \, . \tag{2.10}$$

A commonly-adopted starting point for proving results for queues of $M/G/1$ type, as introduced by Kendall [1951, 1953], is the analysis of a Markov chain embedded just after departure (or service-completion) epochs. Our final preliminary result is a relation between the solutions to the service-start-epoch and the departure-epoch embedded Markov chains in $M/G/1/K$: Let δ_j, $0 \le j \le K-1$, be the proportion (defined similar to (2.7)) of departing *served* customers who leave j other customers behind; then,

$$\delta_0 = \sigma_0 a_0 \, , \tag{2.11}$$

$$\delta_1 = \sigma_0 \, (1 - a_0) \, , \tag{2.12}$$

and

$$\delta_j = \sigma_{j-1}, \qquad 2 \le j \le K-1 \, . \tag{2.13}$$

We will prove (2.11), (2.12), and (2.13), which appear to be new, in Section 3.1, using a direct sample-path argument.

We are now ready for the statement of our main result, which gives transform-free formulas for α_0^+ and $\alpha_{ij}^+(x)$ in terms of σ_0, σ_1, $\cdots$, σ_{K-2} and $\nu_j(x)$:

Theorem 1 *The state distribution of* $\mathbf{Z}^+$ *as seen by a randomly-selected arriving customer is, w.p.1, given by*

$$\alpha_0^+ = (\sigma_0 a_0 + \rho)^{-1} \sigma_0 a_0 \tag{2.14}$$

and

$$\alpha_{ij}^+(x) = (1 - \alpha_0^+) \, \sigma_i \, \nu_j(x) \, , \qquad 0 \le i \le K-2, \, 0 \le j < \infty, \ \textit{and} \ x \ge 0 \, . \tag{2.15}$$

Furthermore, let η *be the proportion of arriving customers who actually enter the system (and receive service); then, w.p.1,*

$$\eta = (\sigma_0 a_0 + \rho)^{-1} \, . \tag{2.16}$$

The explicit term-by-term interpretations for the formulas (2.14) and (2.15) are as follows: Consider a randomly-selected arriving customer, whom we shall denote by C_r (see our discussion after (1.4) and (1.5)). Then, in (2.14), $(\sigma_0 a_0 + \rho)^{-1}$ (see (2.16)) is the probability for C_r to enter the system; and $\sigma_0 a_0$ equals the conditional probability for C_r to also find the system empty. And in (2.15), $(1 - \alpha_0^+)$ is the probability for C_r to be blocked; and σ_i and $\nu_j(x)$ are the conditional probabilities for C_r to also find, in the service interval "interrupted" by the arrival of C_r, $Q_s(\cdot) = i$ and, respectively, $Q_a(\cdot) = j$ and $R(\cdot) \leq x$. Moreover, the most important assertion in (2.15) is that the status of $Q_s(\cdot)$ and the status of $(Q_a(\cdot), R(\cdot))$ are, *from the viewpoint of the blocked customer* C_r, "conditionally independent". Our proof of Theorem 1, which rigorously establishes all of the above interpretations, will be given in Section 3.2.

We now turn our attention to the process $\mathbf{Z}$. At a given time epoch t, we say that $\mathbf{Z}$ is in state Θ_0 if $L(t) = 0$; and that it is in state $\Theta_j(x)$, where $1 \leq j \leq K$ and $x \geq 0$, if $L(t) = j$ and $R(t) \leq x$. Similar to (1.4) and (1.5), we define α_0 and $\alpha_j(x)$, $1 \leq j \leq K$ and $x \geq 0$, as the proportions of customers (including those who do not enter) who, on arrival, find $\mathbf{Z}$ in states Θ_0 and $\Theta_j(x)$, respectively. Then, it follows immediately from (1.3) that

$$\alpha_0 = \alpha_0^+ \,, \tag{2.17}$$

$$\alpha_j(x) = \sum_{i=0}^{j-1} \alpha^+_{i,\,j-1-i}(x) \,, \qquad 1 \leq j \leq K-1 \text{ and } x \geq 0 \,, \tag{2.18}$$

and

$$\alpha_K(x) = \sum_{j=K}^{\infty} \sum_{i=0}^{K-2} \alpha^+_{i,\,j-1-i}(x) \,, \qquad x \geq 0 \,; \tag{2.19}$$

and therefore, substitution of (2.14) and (2.15) into the right-hand sides of (2.17), (2.18), and (2.19) yields the following consequence of Theorem 1:

Theorem 2 *The state distribution of* $\mathbf{Z}$ *as seen by a randomly-selected arriving customer is, w.p.1, given by*

$$\alpha_0 = (\sigma_0 a_0 + \rho)^{-1} \sigma_0 a_0 \,, \tag{2.20}$$

$$\alpha_j(x) = (1 - \alpha_0) \sum_{i=0}^{j-1} \sigma_i \, \nu_{j-1-i}(x) \,, \qquad 1 \leq j \leq K-1 \text{ \textit{and} } x \geq 0 \,, \tag{2.21}$$

and

$$\alpha_K(x) = (1 - \alpha_0) \sum_{j=K}^{\infty} \sum_{i=0}^{K-2} \sigma_i \, \nu_{j-1-i}(x) \,, \qquad x \geq 0 \,. \tag{2.22}$$

(See Cohen [1982], pp. 574–575, equations (6.34) and (6.35), for related formulas that are expressed in terms of integrals in the complex plane; also, under the assumption of phase-type services, see Neuts [1981], p. 88, equations (3.2.15) and (3.2.16), for formulas in terms of a rate matrix "R". Our formulas appear to be new.)

Let $\alpha_j \equiv \alpha_j(\infty)$ for $1 \le j \le K$; then α_K, being the proportion of arriving customers that are lost (or the "loss probability"), is of particular interest: Since $1 - \alpha_K = \eta$ by definition, we have immediately from (2.16) that

$$\alpha_K = 1 - (\sigma_0 a_0 + \rho)^{-1}, \tag{2.23}$$

an interesting formula that is due originally to Keilson [1966], p. 197, equation (6.3b) (see also Cooper [1981], p. 237, equation (9.13); and Keilson and Servi [1989], for related recent results). Alternatively, we note that (2.23) can also be obtained from (2.22), by setting $x = \infty$, substituting (2.20) and the easily-established formula $\nu_j(\infty) = (1/\rho) \sum_{i=j+1}^{\infty} a_i$, and some algebra; we omit the details.

By excluding lost customers, we can also easily obtain from Theorem 2 the state distribution of $\mathbf{Z}$ from the viewpoint of a randomly-selected *entering* customer: Denote by η_0 and $\eta_j(x)$, $1 \le j \le K-1$ and $x \ge 0$, the proportions of entering customers who, on arrival, find the process $\mathbf{Z}$ in, respectively, states Θ_0 and $\Theta_j(x)$. Then, since η is the proportion of arriving customers who actually enter, we immediately have $\eta_0 = \alpha_0/\eta$ and $\eta_j(x) = \alpha_j(x)/\eta$; and, upon substitution of (2.20), (2.21), and (2.16), this leads to:

Theorem 3 *The state distribution of* $\mathbf{Z}$ *as seen by a randomly-selected entering customer is, w.p.1, given by*

$$\eta_0 = \sigma_0 a_0 \tag{2.24}$$

and

$$\eta_j(x) = \rho \sum_{i=0}^{j-1} \sigma_i \, \nu_{j-1-i}(x)\,, \qquad 1 \le j \le K-1 \text{ and } x \ge 0\,. \tag{2.25}$$

It is interesting to observe that the right-hand sides of (2.24) and (2.25) depend on K only through the probabilities σ_0, σ_1, $\cdots$, σ_{K-2}, and then in a very direct way; moreover, in (2.25), notice that $\rho\,\nu_j(x) = P\{\mathbf{1}_{\cdot j;\,x}(1) = 1\}$ (see (2.9) and (2.10)). Prompted by these observations, we give an independent sample-path argument for Theorem 3 in Section 3.3.

In the remainder of this section, we state a series of other results that are consequences of our three theorems, and defer their proofs to Section 3.

In Section 3.4, we show that (2.11), (2.12), (2.13), and (2.25) can be used to derive a new transform formula for the state distribution of $\mathbf{Z}$ as seen by a randomly-selected entering customer: Define the probability-generating function

$$\eta^*(z,\, x) \equiv \sum_{j=1}^{K-1} \eta_j(x)\, z^j, \qquad x \ge 0\,, \tag{2.26}$$

and the Laplace-Stieltjes transform

$$\eta^{**}(z,\, s) \equiv \int_0^\infty e^{-sx}\, d_x \eta^*(z,\, x)\,; \tag{2.27}$$

then, we prove that

$$\eta^{**}(z,\, s) = \frac{\lambda}{\lambda - s}\, z \sum_{j=0}^{K-2} z^j \sum_{i=0}^{j} [\sigma_i G^*(s) - \delta_i] \left(\frac{\lambda}{\lambda - s}\right)^{j-i}, \tag{2.28}$$

where G^* denotes the Laplace-Stieltjes transform of the service-time distribution G. Apart from its very interesting form, formula (2.28) is useful, for example, for the calculation of moments and for numerical inversion.

In Section 3.5, we derive from (2.24) and (2.25) a transform-free formula, apparently new, for the distribution of the waiting time (in queue) W of a randomly-selected entering customer:

$$P\{W \le t\} = \sigma_0 a_0 + \sum_{j=0}^{K-2} \int_0^t G^{[j]}(t-x) \int_x^\infty \left(\sum_{i=0}^{j} \sigma_i \frac{[\lambda(y-x)]^{j-i}}{(j-i)!} e^{-\lambda(y-x)} \right) dG(y)\, \lambda\, dx\,, \tag{2.29}$$

where $G^{[n]}$ denotes the n-fold self-convolution of G. Interestingly, by considering an equivalent closed cyclic two-station tandem queue, Lavenberg [1975] (p. 505, equation (8)) has shown that the Laplace-Stieltjes transform of the distribution of W is given by

$$\int_0^\infty e^{-st}\, dP\{W \le t\} = [G^*(s)]^{K-1} \sum_{j=0}^{K-1} \delta_j \left(\frac{\lambda}{\lambda - s}\right)^{K-j} - \delta_0 \left(\frac{s}{\lambda - s}\right) \sum_{j=0}^{K-1} \left(\frac{\lambda G^*(s)}{\lambda - s}\right)^j \tag{2.30}$$

(also, as noted by Lavenberg, see Cohen [1982], p. 577, for another expression in terms of an integral in the complex plane). We also show, in Section 3.5, that (2.30) can be derived from (2.24) and (2.25), via (2.11), (2.12), (2.13), and (2.28); thus (2.29), indeed, is the inversion of (2.30).

In Section 3.6, we compare two $M/G/1$ finite-capacity queues with capacities K and $K+1$: Using superscripts K and $K+1$ to differentiate between the respective models, we prove that

$$\eta_0^{K+1} = \pi^K \eta_0^K\,, \tag{2.31}$$

$$\eta_j^{K+1}(x) = \pi^K \eta_j^K(x)\,, \qquad 1 \le j \le K-1 \text{ and } x \ge 0\,, \tag{2.32}$$

and

$$\eta_K^{K+1}(x) = \rho \left[\pi^K \sum_{i=0}^{K-2} \sigma_i^K \nu_{K-1-i}(x) + (1 - \pi^K)\, \nu_0(x) \right], \qquad x \ge 0\,, \tag{2.33}$$

where

$$\pi^K = \left(1 + a_0^{-1} \sum_{i=0}^{K-2} \sigma_i^K \sum_{j=K}^{\infty} a_{j-i}\right)^{-1}; \tag{2.34}$$

that is, the *joint* distributions of the number of customers present and the remaining service time of the customer (if any) in service as seen by randomly-selected entering customers in the two models are, with the exception of $\eta_K^{K+1}(x)$, *proportional*. This generalizes (to include remaining service times) a striking proportionality result of Keilson [1966] (p. 190, equations (1.1) and (1.2)) for the queue-length distributions (see also Cooper [1981], p. 238, Exercise 21). In particular, our argument, which is based on (2.24) and (2.25), shows that results of this type are inherited from proportionality of the solutions of the service-start-epoch embedded Markov chains for different values of K (see Keilson [1966], Section 5, and Cooper [1981], pp. 235–237, for related discussions).

In Section 3.7, we show (again from (2.24) and (2.25)) that if a randomly-selected entering customer is blocked, then the conditional distribution of the time R needed to complete the service in progress is given by

$$P\{R \le x\} = (1 - \sigma_0 a_0)^{-1} \int_0^{\infty} \left(\sum_{j=0}^{K-2} \sum_{i=0}^{j} \sigma_i \frac{(\lambda y)^{j-i}}{(j-i)!} e^{-\lambda y}\right) [G(y+x) - G(y)]\, \lambda\, dy, \quad x \ge 0, \tag{2.35}$$

another apparently new result (see Mandelbaum and Yechiali [1979] and Krakowski [1989] for related results). Moreover, we show, also in Section 3.7, that if we assume $\rho < 1$ and consider R as a function of K, then

$$\lim_{K \to \infty} P\{R \le x\} = \mu \int_0^x [1 - G(y)]\, dy\,, \qquad x \ge 0\,. \tag{2.36}$$

Notice that the right-hand side of (2.36) is the *equilibrium-excess* (or *forward-recurrence-time*) distribution of a renewal process with interevent-time distribution G; thus, (2.35) agrees with and generalizes a well-known result of Takács [1963] (p. 491, equation (17)) for $M/G/1$.

In Section 3.8, we consider the standard $M/G/1$ queue, that is, let $K = \infty$ and assume $\rho < 1$: Wishart [1961] established that

$$\eta^{**}(z,\, s) = \frac{\lambda z(1-\rho)(1-z)}{G^*(\lambda - \lambda z) - z} \, \frac{G^*(s) - G^*(\lambda - \lambda z)}{\lambda(1-z) - s} \tag{2.37}$$

(see also Cohen [1982], p. 258, equation (4.88)). Subsequently, Takács [1963] rederived (2.37) and observed in a "remark" that it follows from (2.37) "by inversion" (with respect to s) that

$$\eta^*(z,\, x) = \frac{\lambda z(1-\rho)(1-z)}{G^*(\lambda - \lambda z) - z} \int_0^{\infty} e^{-\lambda(1-z)y} [G(y+x) - G(y)]\, dy \tag{2.38}$$

(see also Cohen [1982], p. 257, equation (4.86)). We show that our proof for (2.28) specializes easily to yield formulas (2.37) and (2.38). In the process, we uncover new term-by-term interpretations for the mysterious forms of these classical transform results. (In fact, our (2.25) is the "full" inversion of (2.37), with respect to both z and s.)

Finally, as an aside, we note the following consequence of (2.11), (2.12), and (2.13): The distribution of the number of customers in *queue* (i.e., excluding the customer, if any, in service) immediately after a randomly-selected service-start epoch is identical to that left behind by a randomly-selected departing served customer; that is, $\sigma_0 = \delta_0 + \delta_1$ and $\sigma_j = \delta_{j+1}$ for $1 \leq j \leq K-2$. Moreover, since "arrival" and "departure" averages in discrete-state, skip-free processes coincide (Burke; see Cooper [1981], p. 187, or Gross and Harris [1985], pp. 264–265), implying that $\eta_j = \delta_j$ for $0 \leq j \leq K-1$, where $\eta_j \equiv \eta_j(\infty)$, we see that this identification also carries over to the queue-length distribution as seen by a randomly-selected entering customer. If, in particular, $K = \infty$ and $\rho < 1$, so that $\alpha_j = \eta_j$ for $0 \leq j < \infty$ (since all arriving customers enter), then, the identification extends even further to the arrival queue-length distribution. Thus, formulas (2.11), (2.12), and (2.13) illuminate "the interesting result" of Keilson [1966] (p. 197, equations (6.4) and (6.5)).

3 Proofs

3.1 Proofs of (2.11), (2.12), and (2.13)

We first observe that, with the exceptions of $j = 0$ and 1, every (served) departure leaving j other customers behind also is a service start with $j-1$ customers waiting in the queue, and vice versa. It follows that for every j from 2 to $K-1$, δ_j and σ_{j-1}, being limiting proportions, are identical by definition; and this establishes (2.13).

To establish (2.11) and (2.12), we "split" state 0 of $\mathbf{Q}$ into two, more-detailed states: We say that a service-start epoch with no customer waiting in the queue is of *type 1*, if it is the first in a busy period; and it is of *type 2* otherwise. Let σ_{0_1} and σ_{0_2} be the proportions of service-start epochs of types 1 and 2, respectively; then, we obviously have

$$\sigma_0 = \sigma_{0_1} + \sigma_{0_2} \,. \tag{3.1}$$

Next, observe that for every type-1 service-start epoch, there corresponds exactly one departure (the last one in the ensuing busy period, more precisely) leaving the system empty; and also that every type-2 service-start epoch is identified with a departure leaving one customer behind. Hence, $\delta_0 = \sigma_{0_1}$ and $\delta_1 = \sigma_{0_2}$. We shall, therefore, complete the proof by showing that

$$\sigma_{0_1} = \sigma_0 a_0 \tag{3.2}$$

and

$$\sigma_{0_2} = \sigma_0 \left(1 - a_0\right). \tag{3.3}$$

Since each busy period is initiated by and contains exactly one type-1 service-start epoch, it follows by first considering n busy periods and then letting $n \to \infty$ that

$$\frac{\sigma_{0_2}}{\sigma_{0_1}} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} Y_k \,, \tag{3.4}$$

where Y_k, $k \geq 1$, denotes the number of type-2 service-start epochs within the k^{th} busy period.

Clearly, the sequence $\{Y_k,\, k \geq 1\}$ is i.i.d. To determine the distribution of Y_1, we observe that *every* service-start epoch with no customer waiting in the queue will either, with probability a_0, be the last one in the busy period in progress, or else, with probability $1 - a_0$, "generate" within the same busy period a subsequent type-2 service-start epoch, which occurs when the number of customers in the system drops back to one again. It follows that $P\{Y_1 = j\} = (1 - a_0)^j a_0$ for $j = 0,\, 1,\, \cdots$, with $E(Y_1) = (1 - a_0)/a_0$. Hence, (3.4) simplifies, by the SLLN, to $\sigma_{0_2}/\sigma_{0_1} = (1 - a_0)/a_0$, which, together with (3.1), implies both (3.2) and (3.3); and this completes our proof. □

3.2 Proof of Theorem 1

We begin with (2.14) and (2.16). We will determine α_0^+ and η, together, by solving two equations that relate them: First, we note that the rate at which customers enter the system is, by definition, given by $\lambda\eta$; and that entering customers spend an average of $1/\mu$ units of time in service. Hence, from a standard application of "$H = \lambda G$" (e.g., Heyman and Stidham [1980]), we have that the proportion of *time* the server is busy is given by $\lambda\eta(1/\mu)$, which, since Poisson arrivals see time averages (Wolff [1982]), also equals $1 - \alpha_0^+$, the proportion of arriving customers that are blocked on their arrival. Therefore, w.p.1,

$$1 - \alpha_0^+ = \eta\,\rho \tag{3.5}$$

(a result due originally to Keilson [1966], p. 193, equation (2.7)).

To obtain another relation between α_0^+ and η, we first note that the ratio α_0^+/η, being a "relative proportion", equals the proportion of *entering* customers who find the system empty. Next, observe that, since every entering customer *eventually* initiates exactly one service, there exists a one-to-one correspondence between entering customers and service-start epochs; furthermore, in this correspondence, every entering customer who finds the system empty is identified with a *type-1* service-start epoch (see Section 3.1). It follows that $\alpha_0^+/\eta = \sigma_{0_1}$ w.p.1; and therefore, from (3.2),

$$\frac{\alpha_0^+}{\eta} = \sigma_0 a_0 \,. \tag{3.6}$$

Solving (3.5) and (3.6) for α_0^+ and η now yields both (2.14) and (2.16).

We now turn our attention to (2.15). For $0 \le i \le K-2$, $0 \le j < \infty$, and $x \ge 0$, let

$$\beta_{ij}^{+}(x) \equiv \frac{\alpha_{ij}^{+}(x)}{1-\alpha_0^{+}}\,; \tag{3.7}$$

then, upon comparison of (3.7) and (2.15), we see that we need to prove, w.p.1,

$$\beta_{ij}^{+}(x) = \sigma_i\,\nu_j(x)\,. \tag{3.8}$$

To establish (3.8), observe that, since $(1-\alpha_0^{+})$ is the proportion of arriving customers that are blocked, the ratio on the right-hand side of (3.7) equals the proportion of *blocked* customers who, on arrival, find the process $\mathbf{Z}^{+}$ in state $\Theta_{ij}^{+}(x)$. Therefore, similar to (2.8), we can take the viewpoint of a randomly-selected blocked customer, and rewrite $\beta_{ij}^{+}(x)$ in the following equivalent, explicit form:

$$\beta_{ij}^{+}(x) = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} \mathbf{1}_{ij;\,x}(k)}{\sum_{k=1}^{n} N_k}\,, \tag{3.9}$$

where $\mathbf{1}_{ij;\,x}(k) \equiv \mathbf{1}_{i\cdot}(k)\mathbf{1}_{\cdot j;\,x}(k)$.

To evaluate the right-hand side of (3.9), we further define, for $0 \le i \le K-2$, $0 \le j < \infty$, and $x \ge 0$, the (relative) proportions

$$\beta_{i\cdot}^{+}(\infty) \equiv \frac{\alpha_{i\cdot}^{+}(\infty)}{1-\alpha_0^{+}} \tag{3.10}$$

and

$$\beta_{\cdot j}^{+}(x) \equiv \frac{\alpha_{\cdot j}^{+}(x)}{1-\alpha_0^{+}}\,, \tag{3.11}$$

with interpretations similar to that of $\beta_{ij}^{+}(x)$ but for states $\Theta_{i\cdot}^{+}(\infty)$ and $\Theta_{\cdot j}^{+}(x)$, and with corresponding explicit forms given, respectively, by

$$\beta_{i\cdot}^{+}(\infty) = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} \mathbf{1}_{i\cdot}(k) N_k}{\sum_{k=1}^{n} N_k} \tag{3.12}$$

and (2.8); and we make the key "conditional-independence" claim

$$\beta_{ij}^{+}(x) = \beta_{i\cdot}^{+}(\infty)\,\beta_{\cdot j}^{+}(x)\,. \tag{3.13}$$

Upon comparison of (3.8) and (3.13) and recalling that $\beta_{\cdot j}^{+}(x)$ was evaluated to $\nu_j(x)$ in Section 2 (see (2.10)), we see that our proof will be complete when we establish both (3.13) and

$$\beta_{i\cdot}^{+}(\infty) = \sigma_i\,, \qquad 0 \le i \le K-2\,. \tag{3.14}$$

To prove (3.14), we interpret the *number* of customers who arrive during a service interval that starts with $Q_s(\cdot) = i$ as a "sojourn in state i" in a "discrete-time" (or an ordinal) semi-Markov process. Then, it is easily seen from (3.12) that under this interpretation, $\beta_{i\cdot}^{+}(\infty)$ is the proportion of "time epochs" (or indices) this semi-Markov process spends in state i. Hence, an application of, for example, Theorem 4.8.3 (its proof, more precisely), pp. 131–132, of Ross [1983] yields

$$\beta_{i\cdot}^{+}(\infty) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{i\cdot}(k)\,, \tag{3.15}$$

which, in view of (2.7), establishes (3.14).

Before proceeding to (3.13), we digress to make explicit an important conceptual point regarding (3.12) and (3.15): With respect to states $\Theta_{i\cdot}^{+}(\infty)$ for $0 \le i \le K-2$, the viewpoint of a randomly-selected blocked customer (cf. (3.12)) is identical to that of a randomly-selected service-start epoch (cf. (3.15)); that is, no *bias* is introduced when we "translate" from the former viewpoint to the latter, and vice versa. This unbiasedness is a consequence of the fact that (a) the *distribution* of the number of arrivals in a service interval does not depend on the value of $Q_s(\cdot)$ associated with that interval; and (b) *all* arrivals, if any, in a service interval "see" the same $Q_s(\cdot)$. Therefore, if we allow the possibility of having different types of services, then explicit length-biasing effects will be present when we translate from one of these viewpoints to another, because property (a) will no longer hold. This, indeed, will be the case for our generalizations in Section 4; and the needed "bias corrections" will be discussed there.

We now consider (3.13): Imagine monitoring the status of the process $\mathbf{Z}^{+}$ from the viewpoint of a randomly-selected *blocked* customer C_r; and observe that $Q_a(\cdot)$ and $R(\cdot)$ are determined from information accumulated *during* the service interval interrupted by the arrival of C_r, whereas $Q_s(\cdot)$ depends on the history of the system *prior* to the beginning of that service. Therefore, *intuitively*, (3.13) can be seen as a consequence of both the independent-increments property of Poisson arrivals and the i.i.d.-service-times assumption, *provided* that the arrival epoch of C_r is "at time infinity", so that any effects from the initial condition of $\mathbf{Z}^{+}$ have worn off (imagine, at a *finite* time epoch, the effects on *both* $Q_s(\cdot)$ and $Q_a(\cdot)$ when a "longer" $R(\cdot)$ is selected as a result of sampling bias; see Feller [1971], Chapter I, p. 11, the "waiting-time paradox"). For prudence, we provide below a rigorous formalization of this intuition.

From (3.9) and (3.12), we have (similar to the proof of Lemma 1 in Niu [1988])

$$\frac{\beta_{ij}^{+}(x)}{\beta_{i\cdot}^{+}(\infty)} = \lim_{n\to\infty} \frac{\left(\sum_{k=1}^{n} \mathbf{1}_{ij;\,x}(k)\right) \,/\, \left(\sum_{k=1}^{n} N_k\right)}{\left(\sum_{k=1}^{n} \mathbf{1}_{i\cdot}(k)\, N_k\right) \,/\, \left(\sum_{k=1}^{n} N_k\right)} = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} \mathbf{1}_{ij;\,x}(k)}{\sum_{k=1}^{n} \mathbf{1}_{i\cdot}(k)\, N_k}\,.$$

Observe that $\mathbf{1}_{i\cdot}(k) = 0$ implies $\mathbf{1}_{ij;\,x}(k) = 0$; and that $\mathbf{1}_{i\cdot}(k) = 1$ implies $\mathbf{1}_{ij;\,x}(k) = \mathbf{1}_{\cdot j;\,x}(k)$. Therefore, by skipping terms with $\mathbf{1}_{i\cdot}(k) = 0$ in both the denominator and the numerator

in the last-displayed expression, it is further simplified to

$$\frac{\beta_{ij}^{+}(x)}{\beta_{i\cdot}^{+}(\infty)} = \lim_{m\to\infty} \frac{\sum_{\ell=1}^{m} \mathbf{1}_{ij;\,x}(k_\ell)}{\sum_{\ell=1}^{m} \mathbf{1}_{i\cdot}(k_\ell)\, N_{k_\ell}} = \lim_{m\to\infty} \frac{\sum_{\ell=1}^{m} \mathbf{1}_{\cdot j;\,x}(k_\ell)}{\sum_{\ell=1}^{m} N_{k_\ell}}\,, \tag{3.16}$$

where $\{k_\ell,\, \ell = 1,\, 2,\, \cdots\}$ enumerates the subset of the indices $k = 1,\, 2,\, \cdots$ for which $\mathbf{1}_{i\cdot}(k) = 1$ (that is, if $\mathbf{1}_{i\cdot}(1) = 1,\ \mathbf{1}_{i\cdot}(2) = 0,\ \mathbf{1}_{i\cdot}(3) = 1,\ \cdots$, then $k_1 = 1,\ k_2 = 3,\ \cdots$). Finally, since $\{\mathbf{1}_{\cdot j;\,x}(k),\, k \geq 1\}$ and $\{N_k,\, k \geq 1\}$ are sequences of i.i.d. random variables, the right-hand side of (3.16) is, w.p.1, identical to that of (2.8). This formally establishes (3.13), and our proof of Theorem 1 is complete. □

3.3 Proof of Theorem 3

Recall from the argument leading up to (3.6) that there exists a one-to-one correspondence between entering customers and service-start epochs. Therefore, (2.24) is an immediate consequence of (3.2). It also follows that we can reinterpret $\eta_j(x)$, $1 \leq j \leq K-1$ and $x \geq 0$, as the proportion of *service-start epochs* that "generate" in their respective ensuing service intervals an (entering) arrival finding the system in state $\Theta_j(x)$. Hence, by first considering n (initial) service-start epochs and then letting $n \to \infty$, we have

$$\begin{aligned}\eta_j(x) &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \sum_{i=0}^{j-1} \mathbf{1}_{i,\,j-1-i;\,x}(k) \\ &= \sum_{i=0}^{j-1} \left(\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{i,\,j-1-i;\,x}(k) \right) \end{aligned} \tag{3.17}$$

where the indicator functions are as defined in (3.9). To evaluate the limits on the right-hand side of (3.17), observe that we have

$$\frac{\lim_{n\to\infty} \sum_{k=1}^{n} \mathbf{1}_{i,\,j-1-i;\,x}(k)/n}{\lim_{n\to\infty} \sum_{k=1}^{n} \mathbf{1}_{i\cdot}(k)/n} = \lim_{m\to\infty} \frac{1}{m} \sum_{\ell=1}^{m} \mathbf{1}_{\cdot,\,j-1-i;\,x}(k_\ell)\,, \tag{3.18}$$

where the subsequence $\{k_\ell,\, \ell = 1,\, 2,\, \cdots\}$ is as defined in (3.16). Next, notice that the denominator on the left-hand side of (3.18) equals σ_i (see (2.7)); and the limit on the right-hand side of (3.18) equals $E[\mathbf{1}_{\cdot,\,j-1-i;\,x}(1)]$ (by the SLLN), which evaluates to $\rho\, \nu_{j-1-i}(x)$ (see (2.9) and (2.10)). Therefore, (3.18) becomes, w.p.1,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{i,\,j-1-i;\,x}(k) = \sigma_i\, \rho\, \nu_{j-1-i}(x)\,. \tag{3.19}$$

Substitution of (3.19) into the right-hand side of (3.17) now yields (2.25), completing our proof. □

3.4 Proof of (2.28)

Substituting (2.25) into (2.26) and changing a summation index, we have

$$\eta^*(z,\,x) = \rho\, z \sum_{j=0}^{K-2} z^j \sum_{i=0}^{j} \sigma_i\, \nu_{j-i}(x)\,, \qquad x \geq 0\,, \tag{3.20}$$

which, upon taking transforms (see (2.27)), leads to

$$\eta^{**}(z,\,s) = \rho\, z \sum_{j=0}^{K-2} z^j \sum_{i=0}^{j} \sigma_i\, \nu^*_{j-i}(s)\,, \tag{3.21}$$

where, by definition,

$$\nu^*_j(s) \equiv \int_0^\infty e^{-sx}\, d\nu_j(x)\,, \qquad 0 \leq j < \infty\,. \tag{3.22}$$

To calculate $\nu^*_j(s)$, we first rewrite (2.10) as

$$\nu_j(x) = \int_0^\infty \frac{(\lambda y)^j}{j!} e^{-\lambda y} \frac{G(y+x) - G(y)}{1 - G(y)}\, \mu[1 - G(y)]\, dy\,,$$

which can then be given the following term-by-term probabilistic interpretation: $\mu[1-G(y)]$ is the density function (see the right-hand side of (2.36)) of the age of the service interval interrupted by a randomly-selected *blocked* (not necessarily entering) customer; given that the age at the point of interruption equals y, $[G(y+x)-G(y)]/[1-G(y)]$ is the conditional probability for the excess to not exceed x, and $[(\lambda y)^j/j!]e^{-\lambda y}$ is, independently of the excess, the conditional probability for j customers to arrive during the age. Now, if we condition on the excess (which, by symmetry, also has density $\mu[1-G(y)]$) and look *backward* in time, then a similar probabilistic argument yields

$$\nu_j(x) = \int_0^x \left\{ \int_y^\infty \frac{[\lambda(z-y)]^j}{j!} e^{-\lambda(z-y)}\, d_z \frac{G(z) - G(y)}{1 - G(y)} \right\} \mu[1 - G(y)]\, dy$$

($d_z[G(z) - G(y)]/[1 - G(y)]$ is the conditional "probability" for the age to equal $z - y$), which simplifies to yet another expression that is equivalent to (2.10):

$$\nu_j(x) = \mu \int_0^x \int_y^\infty \frac{[\lambda(z-y)]^j}{j!} e^{-\lambda(z-y)}\, dG(z)\, dy\,. \tag{3.23}$$

Substituting (3.23) into (3.22), we have

$$\nu^*_j(s) = \mu \int_0^\infty e^{-sx} \int_x^\infty \frac{[\lambda(z-x)]^j}{j!} e^{-\lambda(z-x)}\, dG(z)\, dx\,,$$

which, after an interchange of the order of integration and some algebra, leads to

$$\nu_j^*(s) = \frac{1}{\rho}\left(\frac{\lambda}{\lambda - s}\right)^{j+1}\left[G^*(s) - \sum_{k=0}^{j} a_k \left(\frac{\lambda - s}{\lambda}\right)^k\right]. \tag{3.24}$$

Substitution of (3.24) into (3.21) now yields

$$\eta^{**}(z,\, s) = \frac{\lambda}{\lambda - s}\; z \sum_{j=0}^{K-2} z^j \sum_{i=0}^{j} \sigma_i \left(\frac{\lambda}{\lambda - s}\right)^{j-i}\left[G^*(s) - \sum_{k=0}^{j-i} a_k \left(\frac{\lambda - s}{\lambda}\right)^k\right]. \tag{3.25}$$

To evaluate the right-hand side of (3.25) further, observe that, after an interchange of the order of summation, we have

$$\sum_{i=0}^{j} \sigma_i \left(\frac{\lambda}{\lambda - s}\right)^{j-i} \sum_{k=0}^{j-i} a_k \left(\frac{\lambda - s}{\lambda}\right)^k = \sum_{i=0}^{j} \left(\frac{\lambda}{\lambda - s}\right)^{j-i} \sum_{k=0}^{i} \sigma_k\, a_{i-k}\,, \tag{3.26}$$

where, it follows easily from (2.3), (2.4), (2.11), (2.12), and (2.13),

$$\sum_{k=0}^{i} \sigma_k\, a_{i-k} = \delta_i\,, \qquad 0 \le i \le K-2\,. \tag{3.27}$$

Finally, substitution of (3.26) and (3.27) into (3.25) yields (2.28), completing our proof. □

3.5 Proofs of (2.29) and (2.30)

By conditioning on the state of $\mathbf{Z}$ as seen by a randomly-selected entering customer, it immediately follows that

$$P\{W \le t\} = \eta_0 + \sum_{j=1}^{K-1} \int_0^t G^{[j-1]}(t-x)\, d\eta_j(x)\,. \tag{3.28}$$

Substituting (2.24) and (2.25) into (3.28) and changing a summation index, we have

$$P\{W \le t\} = \sigma_0 a_0 + \sum_{j=0}^{K-2} \int_0^t G^{[j]}(t-x)\, \rho \sum_{i=0}^{j} \sigma_i\, d\nu_{j-i}(x)\,,$$

which, upon substitution of (3.23), rearranges to (2.29).

To derive (2.30), observe that we have, from (3.28),

$$\begin{aligned}\int_0^\infty e^{-st}\, dP\{W \le t\} &= \eta_0 + \int_0^\infty e^{-st}\, d_t \sum_{j=1}^{K-1} \int_0^t G^{[j-1]}(t-x)\, d\eta_j(x) \\ &= \eta_0 + \sum_{j=1}^{K-1} \eta_j^*(s)\,[G^*(s)]^{j-1}\,, \end{aligned} \tag{3.29}$$

where, by definition, $\eta_j^*(s) \equiv \int_0^\infty e^{-st} d\eta_j(t)$ for $1 \le j \le K-1$. Comparison of the right-hand side of (3.29) with (2.26) and (2.27) shows that

$$\int_0^\infty e^{-st}\, dP\,\{W \le t\} = \eta_0 + [G^*(s)]^{-1}\, \eta^{**}\,(G^*(s),\, s)\ ,$$

which, after substituting $\eta_0 = \delta_0$ and (2.28), using (2.11), (2.12), and (2.13), and straightforward (lengthy) algebra, simplifies to (2.30). □

3.6 Proofs of (2.31), (2.32), and (2.33)

By equating the up- and down-crossing rates between states $j-1$ and j of $\mathbf{Q}$, we have, equivalent to (2.3), (2.4), and (2.5),

$$\sum_{i=0}^{j-1} \sigma_i^{K+1} \sum_{k=j+1}^{\infty} a_{k-i} = \sigma_j^{K+1} a_0\,, \qquad 1 \le j \le K-1\,, \tag{3.30}$$

and

$$\sum_{i=0}^{j-1} \sigma_i^{K} \sum_{k=j+1}^{\infty} a_{k-i} = \sigma_j^{K} a_0\,, \qquad 1 \le j \le K-2\,. \tag{3.31}$$

Define

$$\pi^K \equiv \frac{\sigma_0^{K+1}}{\sigma_0^K}\,; \tag{3.32}$$

then, it follows immediately from (3.30) and (3.31) by induction on j that

$$\sigma_j^{K+1} = \pi^K \sigma_j^K\,, \qquad 1 \le j \le K-2\,. \tag{3.33}$$

To determine π^K, observe that we have, from (3.32), (3.33), and (2.6),

$$\sum_{i=0}^{K-2} \sigma_i^{K+1} = \pi^K \sum_{i=0}^{K-2} \sigma_i^K = \pi^K\,,$$

implying, by normalization,

$$\sigma_{K-1}^{K+1} = 1 - \pi^K \tag{3.34}$$

(and π^K is necessarily less than 1). Next, substituting $j = K-1$, (3.32), and (3.33) into (3.30), we also have

$$\begin{aligned} \sigma_{K-1}^{K+1} &= a_0^{-1} \sum_{i=0}^{K-2} \sigma_i^{K+1} \sum_{k=K}^{\infty} a_{k-i} \\ &= a_0^{-1} \pi^K \sum_{i=0}^{K-2} \sigma_i^{K} \sum_{k=K}^{\infty} a_{k-i}\,. \end{aligned} \tag{3.35}$$

Finally, equating the right-hand sides of (3.34) and (3.35) leads to (2.34).

We now show that (3.32) and (3.33) carry over to the averages $\eta_j(x)$ for different values of K. First, we observe that (2.31) is an immediate consequence of (2.24) (that is, $\eta_0^{K+1} = \sigma_0^{K+1} a_0$ and $\eta_0^K = \sigma_0^K a_0$) and (3.32). Next, from (2.25), (3.32), and (3.33), we have, for $1 \le j \le K-1$ and $x \ge 0$,

$$\begin{aligned} \eta_j^{K+1}(x) &= \rho \sum_{i=0}^{j-1} \sigma_i^{K+1} \nu_{j-1-i}(x) \\ &= \pi^K \rho \sum_{i=0}^{j-1} \sigma_i^K \nu_{j-1-i}(x) \\ &= \pi^K \eta_j^K(x)\,, \end{aligned}$$

establishing (2.32). Finally, with $j = K$ in (2.25), we have

$$\begin{aligned} \eta_K^{K+1}(x) &= \rho \sum_{i=0}^{K-1} \sigma_i^{K+1} \nu_{K-1-i}(x) \\ &= \pi^K \rho \sum_{i=0}^{K-2} \sigma_i^K \nu_{K-1-i}(x) + \rho\, \sigma_{K-1}^{K+1}\, \nu_0(x)\,, \end{aligned}$$

which, upon substitution of (3.34), rearranges to (2.33), completing our proof. □

3.7 Proofs of (2.35) and (2.36)

Since $1 - \eta_0$ is the proportion of entering customers that are blocked on their arrival, we have

$$P\{R \le x\} = (1-\eta_0)^{-1} \sum_{j=1}^{K-1} \eta_j(x)\,. \tag{3.36}$$

Substituting (2.24) and (2.25) into (3.36) and changing a summation index, we have

$$P\{R \le x\} = \rho\,(1-\sigma_0 a_0)^{-1} \sum_{j=0}^{K-2} \sum_{i=0}^{j} \sigma_i\, \nu_{j-i}(x)\,,$$

which, upon substitution of (2.10), rearranges to (2.35).

To prove (2.36), we first note that if $\rho < 1$, then $\{\sigma_i^K,\, i = 0,\, 1,\, \cdots,\, K-2\}$ converges in distribution, as $K \to \infty$, to a non-defective distribution $\{\sigma_i^\infty,\, i = 0,\, 1,\, \cdots\}$ (see, e.g., Theorem 2.2, p. 602, of Gibson and Seneta [1987]). Next, the well-known (and easily-shown) result $\alpha_0^\infty = 1 - \rho$ and the relation $\alpha_0^\infty = \eta_0^\infty = \sigma_0^\infty a_0$ together imply $\sigma_0^\infty = (1-\rho)/a_0$. It follows that

$$\lim_{K\to\infty} \sigma_0^K a_0 = 1 - \rho\,. \tag{3.37}$$

Moreover, since $[(\lambda y)^j/j!]e^{-\lambda y}$ for $0 \le j < \infty$ are (Poisson) probabilities, we have that

$$\lim_{K\to\infty} \sum_{j=0}^{K-2} \sum_{i=0}^{j} \sigma_i^K \, \frac{(\lambda y)^{j-i}}{(j-i)!} e^{-\lambda y} = 1\,. \tag{3.38}$$

In light of (3.37) and (3.38), we see that the right-hand side of (2.35) converges to $(1/\rho)\,\int_0^\infty [G(y+x) - G(y)]\,\lambda\,dy$, which easily simplifies to the right-hand side of (2.36); and our proof is complete. □

3.8 Proofs of (2.37) and (2.38)

We begin with Takács's inversion (2.38). Letting $K \to \infty$ in (3.20) and noting that

$$\sum_{i=0}^{j} \rho_i \, \nu_{j-i}(x)$$

becomes (similar to (3.38)) a *convolution* when the range of j is extended to infinity, we have the following *factorization* (or *decomposition*) result:

$$\eta^*(z,\,x) = \rho\, z\, \sigma^*(z)\, \nu^*(z,\,x)\,, \tag{3.39}$$

where ρ (less than 1) is the proportion of arrivals finding the server busy, z is the p.g.f. (probability-generating function) of 1 (accounting for the customer in service; see (1.3)),

$$\sigma^*(z) \equiv \sum_{j=0}^{\infty} \sigma_j \, z^j \tag{3.40}$$

is the p.g.f. of $\{\sigma_i,\, i = 0,\, 1,\, \cdots\}$, and

$$\nu^*(z,\,x) \equiv \sum_{j=0}^{\infty} \nu_j(x)\, z^j, \qquad x \ge 0\,, \tag{3.41}$$

is the p.g.f. of $\{\nu_j(x),\, j = 0,\, 1,\, \cdots\}$.

It immediately follows from (2.10) and (3.41) that

$$\nu^*(z,\,x) = \mu \int_0^\infty e^{-\lambda(1-z)y} [G(y+x) - G(y)]\,dy\,; \tag{3.42}$$

hence we see, upon comparison of (2.38) with (3.39) and (3.42), that (2.38) will follow if

$$\sigma^*(z) = \frac{(1-\rho)(1-z)}{G^*(\lambda - \lambda z) - z}\,. \tag{3.43}$$

To prove (3.43), define the p.g.f.

$$\delta^*(z) \equiv \sum_{j=0}^{\infty} \delta_j \, z^j \tag{3.44}$$

and observe that, from (2.11), (2.12), (2.13), (3.40), and (3.44), we have $\delta^*(z) = \sigma_0 a_0 \, (1 - z) + z \, \sigma^*(z)$ (valid, in fact, for *all* K), which, since $\sigma_0 a_0 = 1 - \rho$ (see (3.37)) when $K = \infty$ and $\rho < 1$, leads to

$$\delta^*(z) = (1 - \rho)(1 - z) + z \, \sigma^*(z) \, . \tag{3.45}$$

For the stable $M/G/1$ queue, it is well known (Kendall [1951, 1953]) that

$$\delta^*(z) = \frac{(1 - \rho)(1 - z) G^*(\lambda - \lambda z)}{G^*(\lambda - \lambda z) - z} \, , \tag{3.46}$$

where $G^*(\lambda - \lambda z)$ is the p.g.f. of $\{a_j, \, j = 0, \, 1, \, \cdots\}$. Substitution of (3.46) into the left-hand side of (3.45) now leads easily to (3.43), completing our proof of (2.38). (Alternatively, (3.43) can also be derived directly from (2.1) with $K = \infty$, along standard lines, similar to, e.g., Takács [1962], pp. 70–72.)

To prove (2.37), we take Laplace-Stieltjes transforms in (3.39) to obtain

$$\eta^{**}(z, \, s) = \rho \, z \, \sigma^*(z) \, \nu^{**}(z, \, s) \, , \tag{3.47}$$

where, by definition, $\nu^{**}(z, \, s) \equiv \int_0^\infty e^{-sx} \, d_x \nu^*(z, \, x)$. After taking the Laplace-Stieltjes transform of the right-hand side of (3.42), it is straightforward to verify (see, e.g., Tilt [1981], p. 138) that

$$\nu^{**}(z, \, s) = \mu \, \frac{G^*(s) - G^*(\lambda - \lambda z)}{\lambda(1 - z) - s} \, . \tag{3.48}$$

Finally, substitution of (3.43) and (3.48) into (3.47) yields (2.37), completing our proof. □

4 Generalizations

The purpose of this section is to show that our proofs in Section 3 actually provide the basis for a formal "calculus" by which transform-free results can be obtained for a host of similar models *in a mechanical manner*. We will outline the analyses for three specific generalizations of the basic $M/G/1/K$ model that allow, respectively, exceptional first services, server vacations, and semi-Markovian services. It will become apparent that our methods also readily apply to the solution of combinations of these as well as other generalizations, such as state-dependent services, Bernoulli feedback, etc.

4.1 Exceptional First Services

Welch [1964] and Avi-Itzhak, Maxwell, and Miller [1965] proposed a very useful generalization of the standard, infinite-capacity $M/G/1$ queue in which the first service in each busy period is allowed to follow a distribution function $\hat{G}$ that is possibly different from G, the service-time distribution for all other "ordinary" services; and they derived a variety of results about queue-length as well as waiting-time distributions, in transform form. Here, we consider the *finite*-capacity version of their model, which does not appear to have been explicitly studied in the literature before, and we show how to obtain transform-free results similar to those of Section 2.

For $1 \le j \le K$ and $x \ge 0$, let $\hat{\alpha}_j(x)$ be the proportion of customers who, on arrival, find that: (a) there are j customers in the system, (b) an "exceptional" service is in progress, and (c) the time needed to complete that service is not greater than x; and let $\alpha_j(x)$ be the corresponding proportion defined by replacing "exceptional" in (b) by "ordinary". In similar ways, we shall also carry over to this model, without further comment, other notations defined for $M/G/1/K$ in Sections 2 and 3, to avoid repetition; that is, the general rule is notations with a "hat" are with respect to either an exceptional service or the exceptional-first-service model.

We first generalize $\mathbf{Q}$ to a corresponding $\hat{\mathbf{Q}}$: The key idea is to judiciously *design* the state space of $\hat{\mathbf{Q}}$ to include "enough information" so that the state of $\hat{\mathbf{Z}}$ (defined similar to (1.1), with service type supplemented) at the arrival epoch of a blocked customer can be determined (similar to (1.2) and (1.3)) once the status of $\hat{\mathbf{Q}}$ is given. With this in mind, it is then clear that $\hat{\mathbf{Q}}$ should have, in addition to $\{0, 1, \cdots, K-2\}$, an extra state $\hat{0}$ to represent an exceptional-service start with (necessarily) no customer waiting in the queue (this is closely related to "splitting state 0" in Section 3.1); and therefore, after modifying (2.1) and (2.2) accordingly (which we omit), we have that the stationary probabilities $\sigma_{\hat{0}}$, $\sigma_0, \cdots, \sigma_{K-2}$ of $\hat{\mathbf{Q}}$ are determined uniquely by the equations

$$\sigma_{\hat{0}} = \sigma_{\hat{0}} \hat{a}_0 + \sigma_0 a_0 \,, \tag{4.1}$$

$$\sigma_j = \sigma_{\hat{0}} \hat{a}_{j+1} + \sum_{i=0}^{j+1} \sigma_i \, a_{j+1-i} \,, \qquad 0 \le j < K-2 \,, \tag{4.2}$$

$$\sigma_{K-2} = \sigma_{\hat{0}} \sum_{j=K-2}^{\infty} \hat{a}_{j+1} + \sum_{i=0}^{K-2} \sigma_i \sum_{j=K-2}^{\infty} a_{j+1-i} \,, \tag{4.3}$$

and the normalization condition

$$\sigma_{\hat{0}} + \sum_{j=0}^{K-2} \sigma_j = 1 \,. \tag{4.4}$$

We next generalize (2.16) and (2.20) (or (2.14)). Observe that an entering customer initiates an exceptional service if and only if the customer finds the system empty; therefore, the proportions of entering customers who receive exceptional and, respectively, ordinary services are given by η_0 and $1-\eta_0$. Moreover, since there is a one-to-one correspondence between entering customers and service-start epochs, we also have $\eta_0 = \sigma_{\hat{0}}$. Hence, the average time spent in service by entering customers is given by $[\sigma_{\hat{0}}(1/\hat{\mu}) + (1-\sigma_{\hat{0}})(1/\mu)]$. Since $\lambda\eta$ is, by definition, the rate at which customers enter the system, an application of both $H = \lambda G$ and PASTA (as in Section 3.2) yields

$$1 - \alpha_0 = \eta\,[\sigma_{\hat{0}}\hat{\rho} + (1-\sigma_{\hat{0}})\rho]\,; \tag{4.5}$$

and this, together with the fact that $\alpha_0/\eta = \eta_0 = \sigma_{\hat{0}}$, leads, after a little bit of algebra, to $\eta = [\sigma_{\hat{0}}(1+\hat{\rho}) + (1-\sigma_{\hat{0}})\rho]^{-1}$ and $\alpha_0 = [\sigma_{\hat{0}}(1+\hat{\rho}) + (1-\sigma_{\hat{0}})\rho]^{-1}\sigma_{\hat{0}}$, generalizing (2.16) and (2.20), respectively.

To generalize (2.21) and (2.22), we now take the viewpoint of a randomly-selected blocked customer. Observe that the *distribution* of the number of arrivals during a service interval depends on whether the service is exceptional or ordinary. Therefore, we need to first generalize (3.13); the idea is to "condition" on the type of the service interval interrupted by the randomly-selected blocked customer: Classify service-start epochs as either exceptional or ordinary, according to the ensuing service type; and denote the set of indices corresponding to the former by Γ, and the latter by Γ^c. Define (similar to (3.9), (3.12), and the right-hand side of (3.16)) the proportions

$$\beta^+_{\hat{0}j}(x) \equiv \lim_{n\to\infty} \frac{\sum_{k\in\Gamma_n} \mathbf{1}_{\hat{0}j;\,x}(k)}{\sum_{k\in\Gamma_n} \hat{N}_k + \sum_{k\in\Gamma_n^c} N_k}\,, \tag{4.6}$$

$$\beta^+_{\hat{0}\cdot}(\infty) \equiv \lim_{n\to\infty} \frac{\sum_{k\in\Gamma_n} \mathbf{1}_{\hat{0}\cdot}(k)\hat{N}_k}{\sum_{k\in\Gamma_n} \hat{N}_k + \sum_{k\in\Gamma_n^c} N_k}\,, \tag{4.7}$$

and

$$\hat{\beta}^+_{\cdot j}(x) \equiv \lim_{n\to\infty} \frac{\sum_{k\in\Gamma_n} \hat{\mathbf{1}}_{\cdot j;\,x}(k)}{\sum_{k\in\Gamma_n} \hat{N}_k}\,, \tag{4.8}$$

where $\Gamma_n \equiv \Gamma \cap \{1,\,2,\,\cdots,\,n\}$ and $\Gamma_n^c \equiv \Gamma^c \cap \{1,\,2,\,\cdots,\,n\}$. Then, since $\mathbf{1}_{\hat{0}j;\,x}(k) \equiv \mathbf{1}_{\hat{0}\cdot}(k)\hat{\mathbf{1}}_{\cdot j;\,x}(k)$, it follows immediately (similar to (3.16)) that, for $0 \le j < \infty$ and $x \ge 0$,

$$\beta^+_{\hat{0}j}(x) = \beta^+_{\hat{0}\cdot}(\infty)\,\hat{\beta}^+_{\cdot j}(x)\,. \tag{4.9}$$

Furthermore, with $\beta^+_{ij}(x)$, $\beta^+_{i\cdot}(\infty)$, and $\beta^+_{\cdot j}(x)$ defined similar to (4.6), (4.7), and (4.8) respectively, we also have, for $0 \le i \le K-2$, $0 \le j < \infty$, and $x \ge 0$,

$$\beta^+_{ij}(x) = \beta^+_{i\cdot}(\infty)\,\beta^+_{\cdot j}(x)\,; \tag{4.10}$$

and (4.9) and (4.10), together, generalize (3.13).

Clearly, we still have $\beta_{\cdot j}^{+}(x) = \nu_j(x)$ w.p.1 (see (2.10)), and similarly $\hat{\beta}_{\cdot j}^{+}(x) = \hat{\nu}_j(x)$. To calculate $\beta_{\hat{0}\cdot}^{+}(\infty)$ and $\beta_{i\cdot}^{+}(\infty)$ for $0 \le i \le K-2$, we again (see our proof of (3.14)) apply Theorem 4.8.3 of Ross [1983] to obtain

$$\beta_{\hat{0}\cdot}^{+}(\infty) = \frac{\sigma_{\hat{0}} E(\hat{N}_1)}{\sigma_{\hat{0}} E(\hat{N}_1) + (1 - \sigma_{\hat{0}}) E(N_1)} \tag{4.11}$$

and

$$\beta_{i\cdot}^{+}(\infty) = \frac{\sigma_i E(N_1)}{\sigma_{\hat{0}} E(\hat{N}_1) + (1 - \sigma_{\hat{0}}) E(N_1)}\,, \tag{4.12}$$

where $E(\hat{N}_1) = \hat{\rho}$ and $E(N_1) = \rho$; thus, indeed, bias corrections are needed when we translate from the viewpoint of a randomly-selected service-start epoch to that of a blocked customer (see the paragraph after (3.15)).

Finally, after substituting (4.5), (4.9), (4.10), (4.11), and (4.12) into the exceptional-first-service versions of (2.17), (2.18), (2.19), and (3.7) (whose explicit statements we omit) and simplifying, we have, for $x \ge 0$,

$$\hat{\alpha}_j(x) = \eta\, \hat{\rho}\, \sigma_{\hat{0}}\, \hat{\nu}_{j-1}(x)\,, \qquad 1 \le j \le K-1\,, \tag{4.13}$$

$$\alpha_j(x) = \eta\, \rho \sum_{i=0}^{j-1} \sigma_i\, \nu_{j-1-i}(x)\,, \qquad 1 \le j \le K-1\,, \tag{4.14}$$

$$\hat{\alpha}_K(x) = \eta\, \hat{\rho}\, \sigma_{\hat{0}} \sum_{j=K}^{\infty} \hat{\nu}_{j-1}(x)\,, \tag{4.15}$$

and

$$\alpha_K(x) = \eta\, \rho \sum_{j=K}^{\infty} \sum_{i=0}^{K-2} \sigma_i\, \nu_{j-1-i}(x)\,; \tag{4.16}$$

and this generalizes (2.21) and (2.22).

All other results in Section 2 can also be generalized mechanically. For example, after dividing the right-hand sides of (4.13) and (4.14) by η to generalize (2.25) (which can also be proved directly by an argument similar to that in Section 3.3) and noting again that $\eta_0 = \sigma_{\hat{0}}$, it is easily seen that (2.29) generalizes, in form, to

$$\begin{aligned} P\{W \le t\} \;\; = \;\; & \sigma_{\hat{0}} + \sum_{j=0}^{K-2} \int_0^t G^{[j]}(t-x) \int_x^{\infty} \sigma_{\hat{0}} \frac{[\lambda(y-x)]^j}{j!} e^{-\lambda(y-x)}\, d\hat{G}(y)\, \lambda\, dx \\ & + \sum_{j=0}^{K-2} \int_0^t G^{[j]}(t-x) \int_x^{\infty} \left(\sum_{i=0}^{j} \sigma_i \frac{[\lambda(y-x)]^{j-i}}{(j-i)!} e^{-\lambda(y-x)} \right) dG(y)\, \lambda\, dx\,. \end{aligned}$$

4.2 Server Vacations with Exhaustive-Service Discipline

In this generalization, the server takes a vacation immediately after the completion of a busy period. Upon returning from a vacation, the server will either start another vacation if there is no waiting customer, or else begin service and continue serving until there are no more customers in the system. The successive vacation durations are i.i.d. random variables following distribution function $\hat{G}$, and are independent of the arrival and service processes. This finite-capacity exhaustive-service vacation model has been studied previously by Courtois [1980] and Lee [1984]. (For a comprehensive survey of the extensive literature on this and other variations of such vacation models, see Doshi [1990].) Our purpose here, as in Section 4.1, is to outline transform-free results for this particular version of vacation models (other variations can also be analyzed similarly).

Our arguments will be based on the following standard reinterpretation (Keilson [1966], Section 7.1; Doshi [1986], p. 36, last paragraph): The time needed to complete each vacation is conceptually equivalent to an "exceptional" service taken up by the server. With this reinterpretation, the methods and results of Section 4.1 directly apply here, after incorporating the fact that although the server is never "idle", exceptional services (since they are actually vacations) do not contribute to the total customer count. In particular, we note that, since the number of waiting positions during an exceptional-service interval equals K (as opposed to $K-1$, for an ordinary-service interval), a maximum of $K-1$ customers (cf. (2.1)) can be waiting in the queue immediately after the subsequent ordinary-service-start epoch (if any); and hence, the state space of $\hat{\mathbf{Q}}$ should have, in addition to $\{\hat{0}, 0, 1, \cdots, K-2\}$, an extra state $K-1$.

In light of the above discussion, we need to replace (4.3) by

$$\sigma_{K-2} = \sigma_{\hat{0}} \hat{a}_{K-1} + \sum_{i=0}^{K-1} \sigma_i \sum_{j=K-2}^{\infty} a_{j+1-i}$$

and $\sigma_{K-1} = \sigma_{\hat{0}} \sum_{j=K}^{\infty} \hat{a}_j$, while making no changes in (4.1) and (4.2); and we need to add $\hat{\alpha}_0(x)$ to $\hat{\alpha}_j(x)$ and $\alpha_j(x)$ for $1 \le j \le K$, while retaining their meanings in Section 4.1. The expressions in (4.11) and (4.12) also carry over here, after extending the range of i for $\beta_{i\cdot}^{+}(\infty)$ to include $i = K-1$.

Observe that although the server is never idle, the proportion of customers who, on arrival, find the system not holding any (waiting) *ordinary* (or "real") customers is given by $\hat{\alpha}_0(\infty)$; and this, a generalization of (2.20), will be a specialization of (4.18) below. To generalize (2.16), we denote by B the proportion of time the server is busy serving *ordinary* customers, and compute it in two different ways: First, from Theorem 4.8.3 of Ross [1983] (in *continuous time* now), we have that, w.p.1, $B = [(1-\sigma_{\hat{0}})(1/\mu)]\,[\sigma_{\hat{0}}(1/\hat{\mu}) + (1-\sigma_{\hat{0}})(1/\mu)]^{-1}$. Next, from $H = \lambda G$, we also have $B = \lambda\eta(1/\mu)$. Equating these two expressions and

rearranging then yields

$$\eta = (1 - \sigma_{\hat{0}}) \left[\sigma_{\hat{0}}\hat{\rho} + (1 - \sigma_{\hat{0}})\rho\right]^{-1} . \tag{4.17}$$

Since *all* arriving customers are "blocked", it follows immediately (similar to (2.18) and (2.19)) from (4.9) and (4.10) (with $i = K - 1$ included in (4.10)) that, for $x \geq 0$,

$$\hat{\alpha}_j(x) = \beta^+_{\hat{0}\cdot}(\infty)\, \hat{\nu}_j(x) , \qquad 0 \leq j \leq K - 1 , \tag{4.18}$$

$$\alpha_j(x) = \sum_{i=0}^{j-1} \beta^+_{i\cdot}(\infty)\, \nu_{j-1-i}(x) , \qquad 1 \leq j \leq K - 1 , \tag{4.19}$$

$$\hat{\alpha}_K(x) = \beta^+_{\hat{0}\cdot}(\infty) \sum_{j=K}^{\infty} \hat{\nu}_j(x) , \tag{4.20}$$

and

$$\alpha_K(x) = \sum_{j=K}^{\infty} \sum_{i=0}^{K-1} \beta^+_{i\cdot}(\infty)\, \nu_{j-1-i}(x) ; \tag{4.21}$$

and these generalize (2.21) and (2.22). (From the above results, we see that stochastic decomposition results such as those discussed in, e.g., Fuhrmann and Cooper [1985] do *not* hold when the queue capacity is *finite*; however, proportionality results similar to (2.31), (2.32), (2.33), and (2.34) can be exploited to facilitate computations.)

Finally, dividing the right-hand sides of (4.18) and (4.19) by η and substituting (4.17), (4.11), and (4.12) yields immediately a generalization of (2.25), which leads to

$$\begin{aligned} P\{W \leq t\} \;=\; & (1 - \sigma_{\hat{0}})^{-1} \left\{ \sum_{j=0}^{K-1} \int_0^t G^{[j]}(t - x) \int_x^{\infty} \sigma_{\hat{0}} \frac{[\lambda(y - x)]^j}{j!} e^{-\lambda(y-x)}\, d\hat{G}(y)\, \lambda\, dx \right. \\ & \left. + \sum_{j=0}^{K-2} \int_0^t G^{[j]}(t - x) \int_x^{\infty} \left(\sum_{i=0}^{j} \sigma_i \frac{[\lambda(y - x)]^{j-i}}{(j - i)!} e^{-\lambda(y-x)} \right) dG(y)\, \lambda\, dx \right\} \end{aligned}$$

(the factor $(1 - \sigma_{\hat{0}})^{-1}$ reflects the fact that not all service-start epochs are "real"), another generalization of (2.29).

4.3 Semi-Markovian Services

In this generalization, we assume that: (a) there are M (possibly infinite) different types of services; (b) services of type m, $1 \leq m \leq M$, follow distribution G_m, with mean $1/\mu_m$; and (c) a type-m service will be followed by a type-n service with probability p_{mn} that is taken from a given $M \times M$ transition-probability matrix $\mathbf{P}$. The *infinite*-capacity version of this model has been studied previously by Çinlar [1967] and Neuts [1966, 1977, 1986],

among others. In this section, we, again, describe how to obtain transform-free results; and we will be very brief.

We shall append an extra subscript (separated by a semicolon) in our notation to differentiate between different types of services, as illustrated by: For $1 \le j \le K$, $1 \le m \le M$, and $x \ge 0$, let $\alpha_{j;\,m}(x)$ be the proportion of customers who, on arrival, find that (a) there are j customers in the system, (b) a type-m service is in progress, and (c) the time needed to complete that service is not greater than x; and for $1 \le m \le M$, let $\alpha_{0;\,m}$ be the proportion of customers who, on arrival, find that the system is empty and the *ensuing* service will be of type m. Other notations in Sections 2 and 3 are also carried over this way, without further comment.

The state space of $\mathbf{Q}$ is designed to be the set $\{(i,\, m) :\ i = \hat{0},\, 0,\, 1,\, \cdots,\, K-2$ and $1 \le m \le M\}$ where the first component i is interpreted as the number of customers in queue, with the additional stipulation that the service is the first one in a busy period if and only if $i = \hat{0}$, and the second component m is interpreted as the ensuing service type. The stationary probabilities $\sigma_{i;\,m}$ for $i = \hat{0},\ \ 0,\, 1,\, \cdots,\, K-2$ and $1 \le m \le M$ of $\mathbf{Q}$ are determined uniquely by solving a system of linear equations (which we omit) similar to (2.3), (2.4), (2.5), and (2.6) (or (4.1), (4.2), (4.3), and (4.4)).

Observe that a proportion $\sigma_{\hat{0};\,m} + \sum_{i=0}^{K-2} \sigma_{i;\,m}$ of entering customers eventually receive a type-m service. Therefore, from $H = \lambda G$ and PASTA, we have that the proportion of arriving customers that are blocked on their arrival is given by

$$1 - \sum_{m=1}^{M} \alpha_{0;\,m} = \lambda\, \eta \left[\sum_{m=1}^{M} \left(\sigma_{\hat{0};\,m} + \sum_{i=0}^{K-2} \sigma_{i;\,m} \right) \mu_m^{-1} \right], \tag{4.22}$$

where $\eta \equiv (1 - \sum_{m=1}^{M} \alpha_{K;\,m})$. Since every "$(\hat{0},\, m)$" service-start epoch is identified with the arrival epoch of an entering customer finding the system empty *and* initiating a type-m service, we also have $\eta_{0;\,m} = \sigma_{\hat{0};\,m}$ for $1 \le m \le M$, which, together with the fact $\eta_{0;\,m} = \alpha_{0;\,m}/\eta$, implies that

$$\alpha_{0;\,m} = \eta\, \sigma_{\hat{0};\,m}\,. \tag{4.23}$$

Substituting (4.23) into (4.22) and solving for η now yields

$$\eta = \left[\sum_{m=1}^{M} \sigma_{\hat{0};\,m}(1 + \rho_m) + \sum_{i=0}^{K-2} \sigma_{i;\,m}\, \rho_m \right]^{-1},$$

and hence also a formula for $\alpha_{0;\,m}$, via (4.23); this generalizes both (2.16) and (2.20).

By "conditioning" on the service type, formulas (4.9) and (4.10) extend immediately to this model. Therefore, we have, for $1 \le j \le K-1$, $1 \le m \le M$, and $x \ge 0$,

$$\alpha_{j;\,m}(x) = \left(1 - \sum_{m=1}^{M} \alpha_{0;\,m} \right) \left(\beta^{+}_{\hat{0}\cdot;\,m}(\infty)\, \nu_{j-1;\,m}(x) + \sum_{i=0}^{j-1} \beta^{+}_{i\cdot;\,m}(\infty)\, \nu_{j-1-i;\,m}(x) \right)$$

and, for $1 \le m \le M$ and $x \ge 0$,

$$\alpha_{K;\,m}(x) = \left(1 - \sum_{m=1}^{M} \alpha_{0;\,m}\right) \sum_{j=K}^{\infty} \left(\beta^{+}_{\hat{0}\cdot;\,m}(\infty)\,\nu_{j-1;\,m}(x) + \sum_{i=0}^{K-2} \beta^{+}_{i\cdot;\,m}(\infty)\,\nu_{j-1-i;\,m}(x)\right) ;$$

and, since the proportion of blocked customers who find, on arrival, the system in state $\Theta^{+}_{i\cdot;\,m}(\infty)$, $i = \hat{0}$, 0, 1, $\cdots$, $K-2$ and $1 \le m \le M$, is given (similar to (4.11) and (4.12)) by

$$\beta^{+}_{i\cdot;\,m}(\infty) = \sigma_{i;\,m} E(N_m) \left\{\sum_{m=1}^{M} \left(\sigma_{\hat{0};\,m} + \sum_{i=0}^{K-2} \sigma_{i;\,m}\right) E(N_m)\right\}^{-1}, \tag{4.24}$$

where $E(N_m) = \lambda/\mu_m \equiv \rho_m$, these expressions simplify, upon substitution of (4.22) and (4.24), to

$$\alpha_{j;\,m}(x) = \eta\,\rho_m \left(\sigma_{\hat{0};\,m}\,\nu_{j-1;\,m}(x) + \sum_{i=0}^{j-1} \sigma_{i;\,m}\,\nu_{j-1-i;\,m}(x)\right)$$

and

$$\alpha_{K;\,m}(x) = \eta\,\rho_m \sum_{j=K}^{\infty} \left(\sigma_{\hat{0};\,m}\,\nu_{j-1;\,m}(x) + \sum_{i=0}^{K-2} \sigma_{i;\,m}\,\nu_{j-1-i;\,m}(x)\right),$$

generalizing (2.21) and (2.22).

Finally, we note that with additional notation, it is straightforward to generalize other results in Section 2. We omit the details.

5 Future Work

By monitoring the queue size *and* the ongoing arrival "phase" immediately after service-start epochs, it is possible to obtain similar results for the $E_k/G/1/K$ queue (Takács [1961]; Truslove [1975a,b]; Hokstad [1977]) and, in fact, for the more general $PH/G/1/K$ queue, where PH refers to any appropriately chosen family of phase-type distributions (see Neuts [1981], for example). Basically, all we need is "sufficient exponentiality" in the arrival process so that we can analyze the given model by designing a suitable embedded *discrete*-state Markov chain immediately after service-start epochs. Moreover, the arrival process does not even have to be a renewal process. We have chosen, in this paper, to work with the Poisson-arrival case because of its simplicity and its basic importance. Work is underway to accommodate more general arrival processes; the immediate focus there, since the basic method has already been illustrated here, will be on (but not necessarily limited to) computational issues.

Note Added in Proof A proportionality result related to formulas (2.31), (2.32), and (2.33) has recently been established, independently, by Glasserman and Gong [1991] (*Journal of Applied Probability*, Vol. 28, No. 3, p. 653, Theorem 2), using a sample-path argument. Another related paper that has come to our attention is Blondia [1989] (*Stochastic Models*, Vol. 5, No. 2, 273–294).

References

[1] Abate, J., and Whitt, W. [1992]. The Fourier-Series Method for Inverting Transforms of Probability Distributions. *Queueing Systems*, **10**, 5–88.

[2] Avi-Itzhak, B., Maxwell, W. L., and Miller, L. W. [1965]. Queuing with Alternating Priorities. *Operations Research*, **13**, 306–318.

[3] Çinlar, E. [1967]. Time Dependence of Queues with Semi-Markovian Service. *Journal of Applied Probability*, **4**, 356–364.

[4] Cohen, J. W. [1982]. *The Single Server Queue*, Second Edition. North-Holland, Amsterdam. First Edition, 1969, American Elsevier, New York.

[5] Cooper, R. B. [1981]. *Introduction to Queueing Theory*, Second Edition. North-Holland (Elsevier). First Edition, 1972, Macmillan. Republished, 1990, by CEEPress, The George Washington University, Washington, DC.

[6] Courtois, P. J. [1980]. The $M/G/1$ Finite-Capacity Queue with Delays. *IEEE Transactions on Communications*, **COM 28**, 165–172.

[7] Doshi, B. T. [1986]. Queueing Systems with Vacations—a Survey. *Queueing Systems*, **1**, 29–66.

[8] Doshi, B. T. [1990]. Single-Server Queues with Vacations. *Stochastic Analysis of Computer and Communication Systems* (H. Takagi, ed.). North-Holland (Elsevier).

[9] Feller, W. [1971]. *An Introduction to Probability Theory and Its Applications, Vol. II.* Second Edition, Wiley, New York.

[10] Fuhrmann, S. W., and Cooper, R. B. [1985]. Stochastic Decompositions in the $M/G/1$ Queue with Generalized Vacations. *Operations Research*, **33**, 1117–1129.

[11] Gibson, D., and Seneta, E. [1987]. Augmented Truncations of Infinite Stochastic Matrices. *Journal of Applied Probability*, **24**, 600–608.

[12] Gross, D., and Harris, C. M. [1985]. *Fundamentals of Queueing Theory*, Second Edition. Wiley, New York.

[13] Heyman, D. P., and Stidham, S., Jr. [1980]. The Relation Between Customer and Time Averages in Queues. *Operations Research*, **28**, 983–994.

[14] Hokstad, P. [1977]. Asymptotic Behaviour of the $E_k/G/1$ Queue with Finite Waiting Room. *Journal of Applied Probability*, **14**, 358–366.

[15] Keilson, J. [1966]. The Ergodic Queue Length Distribution for Queueing Systems with Finite Capacity. *Journal of the Royal Statistical Society*, **B 28**, 190–201.

[16] Keilson, J., and Servi, L. D. [1989]. The $M/G/1/K$ Blocking Formula and its Generalizations. *Queueing Systems*, submitted.

[17] Kendall, D. G. [1951]. Some Problems in the Theory of Queues. *Journal of the Royal Statistical Society*, **B 13**, 151–185.

[18] Kendall, D. G. [1953]. Stochastic Processes Occurring in the Theory of Queues and Their Analysis by Means of the Imbedded Markov Chain. *The Annals of Mathematical Statistics*, **24**, 338–354.

[19] Krakowski, M. [1989]. System Size and Remaining Service in $M/G/1$. Technical Report (No. GMU/22474/114), George Mason University, Fairfax, VA.

[20] Lavenberg, S. S. [1975]. The Steady-State Queueing Time Distribution for the $M/G/1$ Finite-Capacity Queue. *Management Science*, **21**, 501–506.

[21] Lee, T. T. [1984]. $M/G/1/N$ Queue with Vacation Time and Exhaustive Service Discipline. *Operations Research*, **32**, 774–784.

[22] Mandelbaum, A., and Yechiali, U. [1979]. The Conditional Residual Service Time in the $M/G/1$ Queue. Technical Report, Tel Aviv University, Tel Aviv.

[23] Neuts, M. F. [1966]. The Single Server Queue with Poisson Input and Semi-Markov Service Times. *Journal of Applied Probability*, **3**, 202–230.

[24] Neuts, M. F. [1977]. Some Explicit Formulas for the Steady-State Behavior of the Queue with Semi-Markovian Service Times. *Advances in Applied Probability*, **9**, 141–157.

[25] Neuts, M. F. [1981]. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. The Johns Hopkins University Press, Baltimore.

[26] Neuts, M. F. [1982]. Explicit Steady-State Solutions to Some Elementary Queueing Models. *Operations Research*, **30**, 480–489.

[27] Neuts, M. F. [1986]. Generalizations of the Pollaczek-Khinchin Integral Equation in the Theory of Queues. *Advances in Applied Probability*, **18**, 952–990.

[28] Niu, S.-C. [1988]. Representing Workloads in $GI/G/1$ Queues through the Preemptive-Resume LIFO Queue Discipline. *Queueing Systems*, **3**, 157–178.

[29] Niu, S.-C., and Cooper, R. B. [1989]. Duality and Other Results for $M/G/1$ and $GI/M/1$ Queues, via a New Ballot Theorem. *Mathematics of Operations Research*, **14**, 281–293.

[30] Niu, S.-C., and Cooper, R. B. [1991]. A Duality Relation for Busy Cycles in $GI/G/1$ Queues. *Queueing Systems*, **8**, 203–209.

[31] Ross, S. M. [1983]. *Stochastic Processes*. Wiley, New York.

[32] Stidham, S., Jr. [1974]. A Last Word on $L = \lambda W$. *Operations Research*, **22**, 417–421.

[33] Takács, L. [1961]. Transient Behavior of a Single Server Queueing Process with Erlang Input. *Transactions of the American Mathematical Society*, **100**, 1–28.

[34] Takács, L. [1962]. *Introduction to the Theory of Queues*. Oxford University Press, New York.

[35] Takács, L. [1963]. Delay Distributions for One Line with Poisson Input, General Holding Times, and Various Orders of Service. *The Bell System Technical Journal*, **43**, 487–453.

[36] Tilt, B. [1981]. *Solutions Manual for Robert B. Cooper's Introduction to Queueing Theory*, Second Edition. North-Holland (Elsevier). Republished, 1990, by CEEPress, The George Washington University, Washington, DC.

[37] Truslove, A. L. [1975a]. Queue Length for the $E_k/G/1$ Queue with Finite Waiting Room. *Advances in Applied Probability*, **7**, 215–226.

[38] Truslove, A. L. [1975b]. The Busy Period of the $E_k/G/1$ Queue with Finite Waiting Room. *Advances in Applied Probability*, **7**, 416–430.

[39] Welch, P. D. [1964]. On a Generalized $M/G/1$ Queueing Process in Which the First Customer of Each Busy Period Receives Exceptional Service. *Operations Research*, **12**, 736–752.

[40] Wishart, D. M. G. [1961]. An Application of Ergodic Theorems in the Theory of Queues. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 2*, University of California Press, 581–592.

[41] Wolff, R. W. [1982]. Poisson Arrivals See Time Averages. *Operations Research*, **30**, 223–231.