

Research Article

The $N \times D$ -BMAP/G/1 Queueing Model: Queue Contents and Delay Analysis

**Bart Steyaert, Joris Walraevens, Dieter Fiems,
and Herwig Bruneel**

*Department of Telecommunications and Information Processing (IR07), Ghent University (UGent),
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium*

Correspondence should be addressed to Bart Steyaert, bs@telin.ugent.be

Received 20 May 2011; Accepted 1 July 2011

Academic Editor: Alexander Pogramsy

Copyright © 2011 Bart Steyaert et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We consider a single-server discrete-time queueing system with N sources, where each source is modelled as a correlated Markovian customer arrival process, and the customer service times are generally distributed. We focus on the analysis of the number of customers in the queue, the amount of work in the queue, and the customer delay. For each of these quantities, we will derive an expression for their steady-state probability generating function, and from these results, we derive closed-form expressions for key performance measures such as their mean value, variance, and tail distribution. A lot of emphasis is put on finding closed-form expressions for these quantities that reduce all numerical calculations to an absolute minimum.

1. Introduction

In this paper, we analyze a queueing model with generally distributed customer service times and a correlated arrival process. This kind of model is, for instance, useful in assessing the performance of a packet-switched communication infrastructure, where messages carried by the network can have a variable transmission time, such as the current Internet, or a dedicated packet-based network carrying video-on-demand (VoD). Multiplexer queues and/or switches and routers in such a network can in general be modelled by means of a discrete-time queueing system, where new packets (i.e., “customers”) are generated by a superposition of individual traffic sources. The size of the packets is proportional to their transmission times, whose distribution depends on the specific application under consideration. Analyzing such a system is mandatory in the design and evaluation of these networks. However, this can be a difficult task because of the time-correlated behaviour that each of the individual sources may exhibit.

To facilitate the queueing analysis, a source is often modelled as a discrete-time Markov modulated arrival process, such as the discrete-time batch Markovian arrival process

(D-BMAP; [1, 2]) with the Markov modulated Bernoulli process (MMBP; [3–5]) and the switched batch bernoulli process (SBBP; [6, 7]) as frequently used special cases. As a consequence, the analysis of a queueing model with a specific SBBP [8] or MMBP [4, 9–11] source model or with a general D-BMAP (e.g., [1, 12–21], and the vast amount of references therein) has become the focus of many research papers. In addition to the characteristics of the arrival process, such queueing models can also be classified by means of the service process. For customer service times equal to one slot, single-server queueing models are considered in [4, 9, 12, 15], while [18, 21] focus on the multiserver case. For the single-server case, geometrically [1, 10] and phase-type [13] distributed customer service times have been considered as well, while these quantities are generally distributed in [8, 11, 14, 17, 19, 20]. In this paper, we investigate the scenario where arrivals are generated by an aggregation of N sources, where each source is characterized by means of a (two-state) D-BMAP, and the sources are assumed to be homogeneous, that is, they are all described by the same stochastic process. The service times on the other hand are assumed to be generally distributed.

Although the queueing model that is studied is not entirely new, the novelty of this contribution lies in the way by which the problem is tackled and solved. A widely used method to deal with correlated arrival processes is the *matrix-analytic* solution technique [22, 23], which provides a broad and generic framework for the analysis of a wide set of queueing systems, both for finite- [4, 9–12, 14, 15, 19] and infinite-capacity [8, 24] queues. This is based on an intelligent matrix representation of the quantities that are needed in these analyses and making full use of the properties that these matrices exhibit. One impediment it may nevertheless sometimes suffer from is the size of the matrices that must be manipulated, which is (roughly speaking) determined by the *state space* of the system. An alternative solution technique, adopted in [1, 16–18], is the so-called *spectral decomposition* solution method; however, the main results that are generated still contain operations that can be computationally demanding, such as Kröneckers sums and products, or inversions, of matrices that can have large dimensions. The approach that we adopt—based on the work presented in [21] for single-slot service times—is related to the latter one, albeit that we follow a different path by invoking as much as possible a representation by means of (joint) *probability generating functions* (pgfs) in our derivations presented in the remainder of this paper. The main advantage lies in the observation that the formulae for the performance indices that we finally obtain, such as the moments of the queue contents, unfinished work, and customer delay, are—to a large extent—expressed *directly* in terms of the system parameters and therefore often lend themselves to an interpretation that is helpful for understanding the mechanisms that determine the queue behaviour. In addition, in order to establish (semi)analytic expressions for the relevant performance measures that pertain to the queue behaviour, a (possibly large) set of *boundary probabilities* has to be calculated numerically. To reduce the numerical work as much as possible, we introduce approximations for these quantities that allow us to compute the aforementioned performance measures without great difficulty. By means of some numerical examples we will demonstrate the efficiency and accurateness of these approximations, as well as the influence of the system parameters on the queue behaviour.

2. System Description

We consider a discrete-time queueing model, that is, a system where the time axis is divided into fixed-length contiguous intervals, referred to as *slots*. The system consists of one single server and an infinite waiting room for customers awaiting service. Customers arrive in

the queue according to a correlated Markovian arrival process, described further on. It is assumed that the service of a customer requires a positive integer number of slots, described by a general service time distribution, and can start (and end) at slot marks only, that is, at the time points between consecutive slots, meaning that the customer service is synchronized with respect to the slot marks. This is not necessarily the case as far as customer arrivals are concerned: customers may enter the queue at any continuous time instant. Nevertheless, due to the synchronous nature of a customer's service, it can commence at the earliest beginning of the slot following its arrival (if it arrives in an empty queue). Therefore, the precise details of the position of the customer arrival instants within a slot are irrelevant, and it suffices to characterize the arrival process by a random variable describing the total number of customer arrivals during a slot. We will now discuss the customer arrival and service processes in more detail.

Consider a buffer model fed by N identical and independent sources generating customers with variable service times. Each source is modelled as a 2-state Markov modulated arrival process, where the state of a source during a slot is represented by S_i , $1 \leq i \leq 2$. The customer arrival process during a slot is completely characterized by the 2×2 matrix

$$\mathbf{Q}(z) = \begin{bmatrix} q_{11}(z) & q_{12}(z) \\ q_{21}(z) & q_{22}(z) \end{bmatrix}. \quad (2.1)$$

State transitions occur at slot boundaries, and let us denote by p_{ij} , $1 \leq i, j \leq 2$, the one-step transition probability $\Pr[S_i \rightarrow S_j]$ at the end of slot $k-1$, where of course $p_{j1} + p_{j2} = 1$, $1 \leq j \leq 2$. The elements $q_{ij}(z)$ of $\mathbf{Q}(z)$ are then given by

$$q_{ij}(z) \hat{=} G_{ij}(z)p_{ij}, \quad (2.2)$$

where $G_{ij}(z)$, $1 \leq i, j \leq 2$, is the probability generating function (pgf) describing the number of customers generated during a slot by a source, given that the source is in state S_j during the tagged slot and was in state S_i during the preceding slot, that is,

$$G_{ij}(z) \hat{=} \mathbf{E} \left[z^{e_k(n)} \mid S_i \rightarrow S_j \text{ at the end of slot } k-1 \right], \quad (2.3)$$

where $e_k(n)$, $1 \leq n \leq N$, represents the number of customer arrivals generated by source n during slot k and $\mathbf{E}[\cdot]$ denotes the expected value of the tagged quantity. For the Markov modulated Bernoulli process (MMBP), the number of customers generated by a source during any slot is either zero or equal to one, meaning that each of the generating functions $G_{ij}(z)$ is a linear function of z , that is,

$$G_{ij}(z) = 1 - g_{ij} + g_{ij}z, \quad (2.4)$$

for some parameters g_{ij} satisfying $0 \leq g_{ij} \leq 1$. Although attention is focused on this specific arrival model in the numerical examples, the analysis is general and can also be applied when the $G_{ij}(z)$'s have a more complex form. Moreover, the notational conventions have been chosen such that the derivations and results that are presented in the subsequent sections can

be readily extended to an L -state D-BMAP as well (see also [20]), $L \geq 2$. We confine ourselves to the case of $L = 2$ in this paper in order to enhance the readability of the derivations that follow hereafter, while their level of complexity is that of arbitrary L .

The aggregate customer arrival process is fully determined, once the probability generating matrix $\mathbf{Q}(z)$ has been specified. Let us define $a_{k,j}$, $1 \leq j \leq 2$, as the number of sources in states S_j during slot k . Note that, due to the fact that the total number of sources equals N , these random variables satisfy

$$\sum_{j=1}^2 a_{k,j} = N. \quad (2.5)$$

In the rest of the paper, we will denote by $\mathbf{x}$ the 2×1 column vector $(x_1, x_2)^T$ (where $(\cdot)^T$ represents the matrix transposition operation), and similarly we write $\mathbf{a}_k = (a_{k,1}, a_{k,2})$. Let us also denote by the 2×1 column vector $\mathbf{B}(\mathbf{x}, z)$ the matrix product

$$\mathbf{B}(\mathbf{x}, z) \triangleq (B_1(\mathbf{x}, z), B_2(\mathbf{x}, z))^T = \mathbf{Q}(z)\mathbf{x}, \quad (2.6)$$

and, for a set of random variables $\mathbf{r} = (r_1, r_2)$, define the joint generating function as

$$\mathbf{E}[\mathbf{x}^{\mathbf{r}}] \triangleq \mathbf{E} \left[\prod_{j=1}^2 x_j^{r_j} \right]. \quad (2.7)$$

Then, with the previous definitions, it is not difficult to show that the joint probability generating function of the random variables denoting the number of customer arrivals and the state of the arrival process during a slot can be written in terms of the state of the arrival process during the previous slot, leading to

$$\mathbf{E}[z^{e_{k+1}} \mathbf{x}^{\mathbf{a}_{k+1}}] = \mathbf{E} \left[z^{e_{k+1}} \prod_{j=1}^2 x_j^{a_{k+1,j}} \right] = \mathbf{E} \left[\prod_{j=1}^2 B_j(\mathbf{x}, z)^{a_{k,j}} \right] = \mathbf{E}[(\mathbf{Q}(z)\mathbf{x})^{\mathbf{a}_k}], \quad (2.8)$$

where e_k represents the total number of customer arrivals during slot k , that is,

$$e_k \triangleq \sum_{n=1}^N e_k(n). \quad (2.9)$$

Equation (2.8) fully describes the number of customers generated during consecutive slots by the N customer sources.

Starting from some initial state, the customer arrival process will evolve to a stochastic equilibrium after a sufficiently long period of time, and we define $A(\mathbf{x})$ as the joint probability generating function of the number of sources in states S_1 and S_2 during an arbitrary slot. If we define σ_1 and σ_2 as the steady-state probability that the Markov state of a source during

an arbitrary slot is S_1 and S_2 , respectively, and σ as the 2×1 vector $(\sigma_1, \sigma_2)^T$, which is the solution of the matrix equation

$$\sigma^T = \sigma^T Q(1), \quad \sigma^T \mathbf{1} = 1, \quad (2.10)$$

leading to

$$\sigma_1 = \frac{p_{21}}{p_{12} + p_{21}}, \quad \sigma_2 = \frac{p_{12}}{p_{12} + p_{21}}, \quad (2.11)$$

where $\mathbf{1}$ is the 2×1 column vector with all elements equal to 1, then $A(\mathbf{x})$ equals

$$A(\mathbf{x}) = \left(\sigma^T \mathbf{x} \right)^N. \quad (2.12)$$

It is further assumed that the service times of consecutive customers that arrive in the queue form a set of independent and identically distributed random variables, denoted by s , with common probability mass function

$$s(l) = \Pr[\text{customer service time} = l \text{ slots}], \quad l \geq 1, \quad (2.13)$$

and corresponding probability generating function

$$S(z) = \mathbb{E}[z^s] = \sum_{l=1}^{\infty} s(l) z^l. \quad (2.14)$$

Obviously, we assume that the service time of a customer is at least one slot.

Due to the infinite-queue capacity, the system will reach a stochastic equilibrium only if the equilibrium condition is satisfied. Defining ρ as the load of the system, this implies that

$$\rho \hat{=} NpL_p < 1 \quad (2.15)$$

must hold, where p denotes the mean number of customer arrivals per slot and per source, and $L_p = S'(1)$ represents the mean customer service time. Taking into account the customer arrival process described in this section, it follows that p can be calculated from

$$p = \frac{d}{dz} \left[\sigma^T Q(z) \mathbf{1} \right]_{z=1} = \sum_{i=1}^2 \sigma_i \sum_{j=1}^2 q'_{ij}(1), \quad (2.16)$$

where primes denote derivatives with respect to the argument.

3. The Joint pgf of the State Vector

3.1. System Equations

We first establish the system equations that control the evolution of the number of customers in the queue during consecutive slots. This is inspired by the work presented in [25], where the case of an i.i.d. (i.e., uncorrelated) customer arrival process was considered and analysed by means of the so-called *supplementary variable* method. Let us therefore define the random variable v_k as the *system contents* at the beginning of slot $k + 1$, which is the number of customers in the queue at the beginning of slot $k + 1$, including the one being served (in case of a nonempty queue). In addition to v_k , we also need information about the amount of time the customer in service still requires at the beginning of slot $k + 1$ before being completely served. We therefore define the random variable h_k as the *residual service time*, which is the number of slots required at the beginning of slot $k + 1$ to complete the service of the customer being served in case of a nonempty queue; when $v_k = 0$, we automatically have $h_k = 0$.

Together with the definitions of the previous section, we can now establish the system equations. We must distinguish between the three following cases.

- (1) $h_k = 0 (\Leftrightarrow v_k = 0)$. When the queue is empty at the beginning of a slot, the number of customers in the queue at the beginning of the next slot equals the number of new arrivals, and we find that

$$v_{k+1} = e_{k+1}. \quad (3.1)$$

If no customers have entered the queue during slot $k + 1$, the residual service time remains zero, otherwise it is equal to the service time of a new customer, which leads to

$$h_{k+1} = \begin{cases} 0 & \text{if } (e_{k+1} = 0), \\ s & \text{if } (e_{k+1} > 0). \end{cases} \quad (3.2)$$

- (2) $h_k = 1$. This implies that the customer in service is completely served at the end of slot $k + 1$:

$$v_{k+1} = v_k + e_{k+1} - 1. \quad (3.3)$$

Similar to the previous case, if no customers have entered the queue during slot $k + 1$ and the customer in service was the only one in the queue, then the queue becomes empty and the residual service time equals zero at the beginning of the next slot, otherwise the service of a new customer commences, that is,

$$h_{k+1} = \begin{cases} 0, & \text{if } (e_{k+1} = 0), (v_k = 1), \\ s, & \text{if } (e_{k+1} > 0) \text{ or } (v_k > 1). \end{cases} \quad (3.4)$$

- (3) $h_k > 1$. In this case, the customer being served receives an extra slot of service without the service being completed, and we obtain

$$\begin{aligned} v_{k+1} &= v_k + e_{k+1}, \\ h_{k+1} &= h_k - 1. \end{aligned} \quad (3.5)$$

3.2. Derivation of a Functional Equation

Clearly, in view of (3.1)–(3.5), we need to keep track of the random variables v_k and h_k in our state description. Consequently, due to the correlated first-order Markov nature of the process controlling the number of customer arrivals during a slot, it becomes clear that the set of random variables $(v_k, h_k, \mathbf{a}_k)$ constitutes a four-dimensional Markov state description of the system at the beginning of consecutive slots; this set of four random variables will be referred to as the state vector. Note that, in view of (2.5), one of the random variables $(a_{k,1}, a_{k,2})$ can be omitted from the state description. However, due to reasons of symmetry, we prefer to maintain both random variables in the state description.

Let us now define the joint probability generating function of $(v_k, h_k, \mathbf{a}_k)$ as

$$P_k(z, y, \mathbf{x}) \triangleq \mathbf{E} \left[z^{v_k} y^{h_k} \mathbf{x}^{\mathbf{a}_k} \right] = \mathbf{E} \left[z^{v_k} y^{h_k} \prod_{j=1}^2 x_j^{a_{k,j}} \right]. \quad (3.6)$$

Let us, for the sake of notation, define a partial generating function of a random variable X if an event A occurs as

$$\mathbf{E} \left[z^X \{A\} \right] \triangleq \mathbf{E} \left[z^X \mid A \right] \Pr[A]. \quad (3.7)$$

From system equations (3.1) and (3.2), taking into account that $h_k = 0 \Leftrightarrow v_k = 0$, we then first of all derive that

$$\begin{aligned} \mathbf{E} \left[z^{v_{k+1}} y^{h_{k+1}} \mathbf{x}^{\mathbf{a}_{k+1}} \{v_k = 0, h_k = 0\} \right] &= \mathbf{E} \left[z^{e_{k+1}} y^s \mathbf{x}^{\mathbf{a}_{k+1}} \{v_k = 0, h_k = 0\} \right] \\ &+ \mathbf{E} \left[(y^0 - y^s) \mathbf{x}^{\mathbf{a}_{k+1}} \{h_k = 0, v_k = 0, e_{k+1} = 0\} \right]. \end{aligned} \quad (3.8)$$

In a similar way, we obtain from (3.3) and (3.4)

$$\begin{aligned} \mathbf{E} \left[z^{v_{k+1}} y^{h_{k+1}} \mathbf{x}^{\mathbf{a}_{k+1}} \{h_k = 1\} \right] &= \mathbf{E} \left[z^{v_k + e_{k+1} - 1} y^s \mathbf{x}^{\mathbf{a}_{k+1}} \{h_k = 1\} \right] \\ &+ \mathbf{E} \left[(y^0 - y^s) \mathbf{x}^{\mathbf{a}_{k+1}} \{h_k = 1, v_k = 1, e_{k+1} = 0\} \right]. \end{aligned} \quad (3.9)$$

Finally, (3.5) can also be translated into a relation between z-transforms, leading to

$$\begin{aligned} \mathbf{E} \left[z^{v_{k+1}} y^{h_{k+1}} \mathbf{x}^{\mathbf{a}_{k+1}} \{h_k > 1\} \right] &= \mathbf{E} \left[z^{v_k + e_{k+1}} y^{h_k - 1} \mathbf{x}^{\mathbf{a}_{k+1}} \right] \\ &- y^{-1} \mathbf{E} \left[z^{e_{k+1}} \mathbf{x}^{\mathbf{a}_{k+1}} \{h_k = 0, v_k = 0\} \right] - \mathbf{E} \left[z^{v_k + e_{k+1}} \mathbf{x}^{\mathbf{a}_{k+1}} \{h_k = 1\} \right]. \end{aligned} \quad (3.10)$$

Summation of (3.8)–(3.10) yields

$$\begin{aligned}
 yP_{k+1}(z, y, \mathbf{x}) &= \mathbb{E}\left[z^{v_k+e_{k+1}} y^{h_k} \mathbf{x}^{\mathbf{a}_{k+1}}\right] + (yS(y) - 1)\mathbb{E}[z^{e_{k+1}} \mathbf{x}^{\mathbf{a}_{k+1}} \{h_k = 0, v_k = 0\}] \\
 &\quad + y \sum_{i=0}^1 (1 - S(y))\mathbb{E}[z^{e_{k+1}} \mathbf{x}^{\mathbf{a}_{k+1}} \{h_k = i, v_k = i, e_{k+1} = 0\}] \\
 &\quad + y(S(y) - z)\mathbb{E}[z^{v_k+e_{k+1}-1} \mathbf{x}^{\mathbf{a}_{k+1}} \{h_k = 1\}].
 \end{aligned} \tag{3.11}$$

Note that, given that $\mathbf{a}_k$ is known, the value of $\mathbf{a}_{k+1}$ (and the number of customer arrivals e_{k+1}) is independent of (v_k, h_k) . Therefore, applying (2.8), the previous relation can be transformed into

$$\begin{aligned}
 yP_{k+1}(z, y, \mathbf{x}) &= P_k(z, y, \mathbf{Q}(z)\mathbf{x}) + y(1 - S(y))\varphi_k(\mathbf{Q}(0)\mathbf{x}) + (yS(y) - 1)\varphi_k(\mathbf{Q}(z)\mathbf{x}) \\
 &\quad + y(1 - S(y))\psi_k(0, \mathbf{Q}(0)\mathbf{x}) + y(S(y) - z)\psi_k(z, \mathbf{Q}(z)\mathbf{x}),
 \end{aligned} \tag{3.12}$$

where

$$\begin{aligned}
 \phi_k(\mathbf{x}) &\triangleq \mathbb{E}[\mathbf{x}^{\mathbf{a}_k} \{u_k = 0, h_k = 0\}] = \mathbb{E}\left[\prod_{j=1}^2 x_j^{a_{k,j}} \{u_k = 0, h_k = 0\}\right], \\
 \varphi_k(z, \mathbf{x}) &\triangleq \mathbb{E}[z^{v_k-1} \mathbf{x}^{\mathbf{a}_k} \{h_k = 1\}] = \mathbb{E}\left[z^{v_k-1} \prod_{j=1}^2 x_j^{a_{k,j}} \{h_k = 1\}\right].
 \end{aligned} \tag{3.13}$$

First of all, due to the foregoing definitions, we have that $P_k(0, y, \mathbf{x}) \equiv P_k(0, 0, \mathbf{x}) = \varphi_k(\mathbf{x})$. If we insert $z = 0$ into (3.12), we derive that

$$\varphi_{k+1}(\mathbf{x}) = \varphi_k(\mathbf{Q}(0)\mathbf{x}) + \psi_k(0, \mathbf{Q}(0)\mathbf{x}). \tag{3.14}$$

Let us now assume that the equilibrium condition (2.15) is satisfied, implying that all the functions that occur in (3.12) evolve to a steady-state limit, which will be reflected in the remainder by suppressing the subscript k (which expressed the time dependence in the previous derivations). Together with the previous relation, we finally obtain the functional equation

$$\begin{aligned}
 yP(z, y, \mathbf{x}) &= P(z, y, \mathbf{Q}(z)\mathbf{x}) + y(1 - S(y))\varphi(\mathbf{x}) + (yS(y) - 1)\varphi(\mathbf{Q}(z)\mathbf{x}) \\
 &\quad + y(S(y) - z)\psi(z, \mathbf{Q}(z)\mathbf{x}).
 \end{aligned} \tag{3.15}$$

Equation (3.15), together with definition (3.13), defines a functional equation for $P(z, y, \mathbf{x})$, the steady-state joint generating function of the random variables $(v_k, h_k, \mathbf{a}_k)$, that is, the system contents and residual service time at the beginning of an arbitrary slot, and the state of the Markovian customer arrival process during the preceding slot.

3.3. Solution of the Functional Equation

In order to solve the functional equation, we follow a similar approach to that in [21]. First, if we substitute consecutive arguments $\mathbf{x} = \mathbf{Q}(z)^i \mathbf{x}'$, $0 \leq i \leq n$, into (3.15), we obtain

$$\begin{aligned} P(z, y, \mathbf{x}) &= y^{-n} P(z, y, \mathbf{Q}(z)^n \mathbf{x}) + (1 - S(y)) \sum_{i=0}^{n-1} y^{-i} \phi(\mathbf{Q}(z)^i \mathbf{x}) \\ &+ (yS(y) - 1) \sum_{i=1}^n y^{-i} \phi(\mathbf{Q}(z)^i \mathbf{x}) + y(S(y) - z) \sum_{i=1}^n y^{-i} \psi(z, \mathbf{Q}(z)^i \mathbf{x}). \end{aligned} \quad (3.16)$$

We now let n approach infinity. Let us therefore express $\mathbf{Q}(z)^n$ in terms of its eigenvalues and eigenvectors as $\mathbf{W}(z)\mathbf{\Lambda}(z)^n\mathbf{U}(z)$. For further details concerning the eigenvalues and eigenvectors of $\mathbf{Q}(z)$, we refer to the appendix. Define the functions $F_{lm}(z)$, $1 \leq l, m \leq N$, as

$$\left(\sum_{j=1}^2 u_{1j}(z) x_j \right)^l \left(\sum_{j=1}^2 u_{2j}(z) x_j \right)^{N-l} \triangleq \sum_{m=0}^N F_{lm}(z) (x_1^m x_2^{N-m}). \quad (3.17)$$

In the above expression, the functions $u_{ij}(z)$ are the components of the column eigenvectors of $\mathbf{Q}(z)$, given by (A.5). One can easily derive that this relation implies that these functions can be written as

$$F_{lm}(z) = \sum_{j=(l+m-N)^+}^{\min\{l,m\}} \binom{l}{j} \binom{N-l}{m-j} u_{11}(z)^j u_{12}(z)^{l-j} u_{21}(z)^{m-j} u_{22}(z)^{N-l-m+j}, \quad (3.18)$$

where $\binom{x}{y} = x!/(y!(x-y)!)$ and $(\cdot)^+ = \max\{0, \cdot\}$. In view of definition (3.17) and defining

$$p_l \triangleq \lim_{k \rightarrow \infty} \Pr[v_k = 0, h_k = 0, a_{k,1} = l, a_{k,2} = N-l], \quad (3.19)$$

for $n \rightarrow \infty$, the first sum in the right-hand side of (3.16) can be written as

$$\begin{aligned} &\sum_{i=0}^{\infty} y^{-i} \phi(\mathbf{Q}(z)^i \mathbf{x}) \\ &= \sum_{i=0}^{\infty} y^{-i} \sum_{l=0}^N \left(\sum_{j=1}^2 u_{1j}(z) \lambda_j(z)^i \mathbf{w}_j(z) \mathbf{x} \right)^l \left(\sum_{j=1}^2 u_{2j}(z) \lambda_j(z)^i \mathbf{w}_j(z) \mathbf{x} \right)^{N-l} p_l, \end{aligned} \quad (3.20)$$

where $\lambda_j(z)$ and $\mathbf{w}_j(z)$ constitute the j th eigenvalue and corresponding row eigenvector of $\mathbf{Q}(z)$. The Perron-Fröbenius (P-F) eigenvalue is denoted by $\lambda_1(z)$ (see the appendix for further details). Due to definition (3.17), this can be transformed into

$$\sum_{i=0}^{\infty} y^{-i} \varphi(\mathbf{Q}(z)^i \mathbf{x}) = \sum_{i=0}^{\infty} y^{-i} \sum_{l=0}^N \sum_{m=0}^N F_{lm}(z) (\lambda_1(z)^i \mathbf{w}_1(z) \mathbf{x})^m (\lambda_2(z)^i \mathbf{w}_2(z) \mathbf{x})^{N-m} p_l. \quad (3.21)$$

Let us now consider values of y and z for which $|\lambda_j(z)| < |y| \leq 1$, $1 \leq j \leq 2$. Such values of y and z exist; this is for instance the case if $|y| = 1$, and $|z| = 1$ and $z \neq 1$ (e.g., see [26]) since

$$|\lambda_1(z)| < 1, \quad |\lambda_2(z)| < 1, \quad \text{for } |z| = 1, \quad z \neq 1, \quad (3.22)$$

unless some special cases for the generating functions $G_{ij}(z)$ defined in (2.3) are considered—for example, the case that the number of customer arrivals per slot is always a multiple of n , $n > 1$ —a situation that falls beyond the scope of this paper. The sum for i in the above expression then converges, and we obtain

$$\sum_{i=0}^{\infty} y^{-i} \varphi(\mathbf{Q}(z)^i \mathbf{x}) = \sum_{m=0}^N \frac{y (\mathbf{w}_1(z) \mathbf{x})^m (\mathbf{w}_2(z) \mathbf{x})^{N-m}}{y - \lambda_1(z)^m \lambda_2(z)^{N-m}} \sum_{l=0}^N F_{lm}(z) p_l. \quad (3.23)$$

In a similar way we can derive that

$$\begin{aligned} \sum_{i=1}^{\infty} y^{-i} \varphi(\mathbf{Q}(z)^i \mathbf{x}) &= \sum_{m=0}^N \frac{(\lambda_1(z) \mathbf{w}_1(z) \mathbf{x})^m (\lambda_2(z) \mathbf{w}_2(z) \mathbf{x})^{N-m}}{y - \lambda_1(z)^m \lambda_2(z)^{N-m}} \sum_{l=0}^N F_{lm}(z) p_l, \\ \sum_{i=1}^{\infty} y^{-i} \varphi(z, \mathbf{Q}(z)^i \mathbf{x}) &= \sum_{m=0}^N \frac{(\lambda_1(z) \mathbf{w}_1(z) \mathbf{x})^m (\lambda_2(z) \mathbf{w}_2(z) \mathbf{x})^{N-m}}{y - \lambda_1(z)^m \lambda_2(z)^{N-m}} \sum_{l=0}^N F_{lm}(z) \varphi_l(z), \end{aligned} \quad (3.24)$$

where

$$\varphi_l(z) \triangleq \lim_{k \rightarrow \infty} \mathbf{E} \left[z^{v_k-1} \{h_k = 1, a_{k,1} = l, a_{k,2} = N-l\} \right]. \quad (3.25)$$

In addition, note that

$$\begin{aligned} \lim_{n \rightarrow \infty} y^{-n} P(z, y, \mathbf{Q}(z)^n \mathbf{x}) &= \sum_{j=0}^{\infty} z^j \sum_{k=0}^{\infty} y^k \sum_{l=0}^N \sum_{m=0}^N F_{lm}(z) \Pr[v = j, h = k, a_1 = l, a_2 = N-l] v \\ &\quad \cdot \lim_{n \rightarrow \infty} y^{-n} (\lambda_1(z)^n \mathbf{w}_1(z) \mathbf{x})^m (\lambda_2(z)^n \mathbf{w}_2(z) \mathbf{x})^{N-m}, \end{aligned} \quad (3.26)$$

which, for $|\lambda_j(z)| < |y| \leq 1$, $1 \leq j \leq 2$, becomes equal to 0. Summarizing, in view of the previous results, (3.16) becomes

$$P(z, y, \mathbf{x}) = \sum_{m=0}^N \frac{(\mathbf{w}_1(z)\mathbf{x})^m (\mathbf{w}_2(z)\mathbf{x})^{N-m}}{y - A_m(z)} \times \sum_{l=0}^N F_{lm}(z) \{ [y(S(y) - z)\psi_l(z) + (yS(y) - 1)p_l] A_m(z) + y(1 - S(y))p_l \}, \quad (3.27)$$

where $A_m(z)$ has been defined as

$$A_m(z) \triangleq \lambda_1(z)^m \lambda_2(z)^{N-m}. \quad (3.28)$$

Expression (3.27) for the joint probability generating function $P(z, y, \mathbf{x})$ still contains the unknown constants p_l , as well as the unknown functions $\psi_l(z)$. Let us therefore, for each value of m , consider values of y for which $y = A_m(z)$, which will be represented by $y_m(z)$. Note that if $|z| \leq 1$, due to (3.22) this solution for y also satisfies $|y_m(z)| \leq 1$, and, for these values of z and y , $P(z, y, \mathbf{x})$ is finite. This means that, if we multiply both hand sides of (3.27) by $y - A_m(z)$ and consider $y = y_m(z)$, then the left-hand side becomes equal to zero, implying that the same must hold for the right-hand side. For each value of $0 \leq m \leq N$, this leads to the following equation:

$$\sum_{l=0}^N F_{lm}(z) \{ A_m(z)(S(A_m(z)) - z)\psi_l(z) + S(A_m(z))(A_m(z) - 1)p_l \} = 0. \quad (3.29)$$

If we insert this equation into (3.27), we finally obtain

$$P(z, y, \mathbf{x}) = \sum_{m=0}^N (\mathbf{w}_1(z)\mathbf{x})^m (\mathbf{w}_2(z)\mathbf{x})^{N-m} \cdot \left\{ 1 + zy \frac{(A_m(z) - 1)(S(y) - S(A_m(z)))}{(y - A_m(z))(z - S(A_m(z)))} \right\} \sum_{l=0}^N F_{lm}(z)p_l, \quad (3.30)$$

This final result for $P(z, y, \mathbf{x})$ expresses the joint generating function of the state of the arrival process during an arbitrary slot, and the queue contents and residual service time at the beginning of the next slot, in terms of the unknown constants p_l . In the next section, it will become clear how these can be calculated or approximated. Also, starting from this result, we will establish expressions for the generating functions of quantities such as the queue contents, unfinished work, and message delay, and their corresponding moments and tail distribution.

3.4. Calculation of the Unknowns p_l

We can calculate these probabilities, defined in (3.19), by expressing that $P(z, y, x)$ is analytic when the complex variables y and z both lie inside the unit circle, that is, $|z| \leq 1$ and $|y| \leq 1$. We have already expressed in the previous section that, for $|y| \leq 1$ and $y = A_m(z)$, $P(z, y, x)$ is finite. In view of expression (3.30), the only remaining potential singularities inside the complex unit disk are those values of z (and y) for which $z = S(A_m(z))$. Indeed, taking into account the definition for $A_m(z)$, then from (3.22) and Rouché's theorem it follows that this equation has exactly one solution inside the complex unit circle. Let us, for each value of $0 \leq m \leq N$, denote this solution by $z_{\{m\}}$, that is,

$$z_{\{m\}} = S(A_m(z_{\{m\}})), \quad |z_{\{m\}}| \leq 1, \quad 0 \leq m \leq N. \quad (3.31)$$

Then, by expressing that $z_{\{m\}}$ must also be a zero of the corresponding numerator in expression (3.30) for $P(z, y, x)$, we obtain

$$\xi_m(z_{\{m\}}) = 0, \quad 0 \leq m \leq N - 1, \quad (3.32)$$

where

$$\xi_m(z) \triangleq \frac{1}{1 - \rho} \sum_{l=0}^N F_{lm}(z) p_l. \quad (3.33)$$

Note that $z_{\{N\}} = 1$ (since $\lambda_1(1) = 1$, see (A.6), implying $A_N(1) = 1$), which leads to no additional equation for the unknowns. Together with the normalization condition, which yields $\xi_N(1) = 1$ (as we will see further on), (3.32) forms a set of $N + 1$ linear equations for the $N + 1$ unknowns p_l that can easily be solved.

However, in deriving numerical results for the performance measures considered further in this paper, we want to avoid all numerical calculations as much as possible. This makes sense in the particular case that N is large, that is, when a high number of sources are multiplexed, since the number of equations to be solved equals $N + 1$. Therefore we also present an approximation for the boundary probabilities p_l . As before, denote by the random variables e and v the number of customer arrivals during a tagged arbitrary slot and the queue contents at the beginning of the following slot; also, $\mathbf{a}$ represents the state of the Markovian arrival process during the tagged slot. We define the steady-state joint probability

$$q_l \triangleq \Pr[a_1 = l, a_2 = N - l, e = 0]. \quad (3.34)$$

Obviously, $v = 0$ implies that there have been no customer arrivals during the tagged slot, that is, $v = 0 \Rightarrow e = 0$. It is therefore clear that the inequality $q_l \geq p_l$ holds. We will show by some numerical examples that approximating the boundary probabilities by

$$p_l \cong \frac{1 - \rho}{\Pr[e = 0]} q_l \quad (3.35)$$

yields excellent approximations for the performance measures of interest. Here, we have taken into account that $\Pr[v = 0] = 1 - \rho$ so that substituting p_l by this formula in (3.30) still yields a normalized generating function.

An additional advantage of the approximation proposed here is that it avoids the explicit calculation of the q_l 's, as well as the functions $F_{lm}(z)$ from (3.18), that occur in expression (3.30) for $P(z, y, x)$. From definition (3.17) and expression (2.8) it is not difficult to show that

$$\sum_{l=0}^N F_{lm}(z) q_l = \binom{N}{m} \left(\sigma^T \mathbf{Q}(0) \mathbf{u}_1(z) \right)^m \left(\sigma^T \mathbf{Q}(0) \mathbf{u}_2(z) \right)^{N-m}, \quad (3.36)$$

where $\mathbf{u}_j(z)$ represents the j th column eigenvector of $\mathbf{Q}(z)$. In view of the previous, this implies that the following approximation may be applied:

$$\xi_m(z) \cong \frac{1}{\Pr[e = 0]} \binom{N}{m} \left(\sigma^T \mathbf{Q}(0) \mathbf{u}_1(z) \right)^m \left(\sigma^T \mathbf{Q}(0) \mathbf{u}_2(z) \right)^{N-m}. \quad (3.37)$$

To conclude, note that it immediately follows from (2.8) that $\Pr[e = 0]$ satisfies

$$\Pr[e = 0] = \left(\sigma^T \mathbf{Q}(0) \mathbf{1} \right)^N. \quad (3.38)$$

4. The Queue Contents

4.1. Derivation of the Probability Generating Function

First of all, due to property (A.2), we have that $\mathbf{w}_j(z) \mathbf{1} = 1$, $1 \leq j \leq 2$. Therefore, inserting $y = 1$ and $x = \mathbf{1}$ into (3.30), we obtain the following expression for the probability generating function $V(z)$ describing the number of customers in the queue at the beginning of an arbitrary slot:

$$V(z) = (1 - \rho) \sum_{m=0}^N \frac{(z - 1) S(A_m(z))}{z - S(A_m(z))} \xi_m(z). \quad (4.1)$$

From (A.7) (see the appendix) and (2.15), it follows that $S'(1) A_N'(1) = N S'(1) \lambda_1'(1) = \rho$, implying that the normalization condition for $V(z)$ indeed requires that $\xi_N(1) = 1$, that is, $\Pr[v = 0] = 1 - \rho$. From (4.1), one can easily derive the performance measures of interest related to the queue contents, as shown next.

4.2. Moments of the Queue Contents

The mean and variance of the queue contents v can be calculated by taking the appropriate derivatives of expression (4.1) for $V(z)$. First, note that (A.9) implies that $F_{lm}(1) = \delta(N - m)$ (where $\delta(n)$ represents the discrete Krönecker-delta function, which equals 1 if n is 0 and 0

otherwise). Since this, in turn, implies that $\xi_m(1) = \delta(N - m)$, and defining $\xi_m(z) \triangleq S(A_m(z))$, we obtain the following expression for mean and variance of the queue contents:

$$\begin{aligned} \mathbf{E}[v] &= \xi'_N(1) + \frac{\xi''_N(1)}{2(1 - \xi'_N(1))} + \xi'_N(1), \\ \mathbf{Var}[v] &= \xi''_N(1) + \xi'_N(1) - \xi'_N(1)^2 + \frac{\xi'''_N(1)}{3(1 - \xi'_N(1))} \\ &\quad + \frac{\xi''_N(1)}{2(1 - \xi'_N(1))} \left(1 + \frac{\xi''_N(1)}{2(1 - \xi'_N(1))} \right) + \xi''_N(1) + \xi'_N(1) - \xi'_N(1)^2 + T'(1), \end{aligned} \quad (4.2)$$

where, with (A.7) and (2.15),

$$\begin{aligned} \xi'_N(1) &= NpL_p = \rho, \\ \xi''_N(1) &= N^2p^2S''(1) + (N-1)\rho p + NL_p\lambda''_1(1), \\ \xi'''_N(1) &= N^3p^3S'''(1) + 3Np \left((N-1)L_p\lambda''_1(1) + NS''(1)\lambda''_1(1) + N(N-1)S''(1)p^2 \right) \\ &\quad + \rho(N-1)(N-2)p^2 + NL_p\lambda'''_1(1). \end{aligned} \quad (4.3)$$

In these expressions, primes denote derivatives with respect to the argument, as before. From definition (3.33) for $\xi_m(z)$ and the observation that $F_{lN}(z) = u_{11}(z)^l u_{12}(z)^{N-l}$, we also find that

$$\begin{aligned} \xi'_N(1) &= \frac{1}{1-\rho} \sum_{l=0}^N (lu'_{11}(1) + (N-l)u'_{21}(1))p_l, \\ \xi''_N &= \frac{1}{1-\rho} \sum_{l=0}^N (lu''_{11}(1) + 2l(N-l)u'_{11}(1)u'_{21}(1) + (N-l)u''_{21}(1) \\ &\quad + l(l-1)u'_{11}(1)^2 + (N-l)(N-l-1)u'_{21}(1)^2)p_l. \end{aligned} \quad (4.4)$$

If we choose to use approximation (3.37) for $\xi_m(z)$ in order to avoid the numerical calculation of the p_l 's, this becomes

$$\begin{aligned} \xi'_N(1) &\cong N \left(\sigma^T \mathbf{Q}(0) \mathbf{u}'_1(1) \right) \left(\sigma^T \mathbf{Q}(0) \mathbf{1} \right)^{-1}, \\ \xi''_N(1) &\cong N \left(\sigma^T \mathbf{Q}(0) \mathbf{u}''_1(1) \right) \left(\sigma^T \mathbf{Q}(0) \mathbf{1} \right)^{-1} + N(N-1) \left(\sigma^T \mathbf{Q}(0) \mathbf{u}'_1(1) \right)^2 \left(\sigma^T \mathbf{Q}(0) \mathbf{1} \right)^{-2}. \end{aligned} \quad (4.5)$$

In addition, the term $T'(1)$ in the expression for $\mathbf{Var}[v]$ is given by

$$T'(1) = (1-\rho) \sum_{m=0}^{N-1} \frac{A_m(1)}{1-A_m(1)} \xi'_m(1), \quad (4.6)$$

which can be considered as the first derivative with respect to z of $T(z)$ for $z = 1$, where, using (3.22) and the definition for $\xi_m(z)$ and $A_m(z)$, $T(z)$ satisfies

$$\begin{aligned} T(z) &= \frac{1}{1-\rho} \sum_{m=0}^{N-1} \frac{A_m(1)}{1-A_m(1)} \xi_m(z) = \sum_{l=0}^N \sum_{k=1}^{\infty} \left\{ -F_{lN}(z) + \sum_{m=0}^N A_m(1)^k F_{lm}(z) \right\} p_l \\ &= \sum_{l=0}^N \sum_{k=1}^{\infty} \left\{ -u_{11}(z)^l u_{21}(z)^{N-l} + \left(\sum_{j=1}^2 u_{1j}(z) \lambda_j(1)^k \right)^l \left(\sum_{j=1}^2 u_{2j}(z) \lambda_j(1)^k \right)^{N-l} \right\} p_l, \end{aligned} \quad (4.7)$$

where we have also taken into account that $\lambda_1(1) = A_N(1) = 1$. Taking the first-order derivative, we thus obtain, keeping in mind that $u_{j2}(1) = 0$ and $u'_{j2}(1) = -u'_{j1}(1)$, $1 \leq j \leq 2$ (see (A.9)–(A.10))

$$T'(1) = -(1-\rho) \frac{\lambda_2(1)}{1-\lambda_2(1)} \xi'_N(1). \quad (4.8)$$

By using expressions (4.1)–(4.8), we have now expressed the moments of the queue contents exclusively in terms of the system parameters and the derivatives of the P-F eigenvalue and eigenvector given in (A.7)–(A.11).

4.3. Tail Distribution of the Queue Contents

It has been observed in many cases (see, e.g., [16, 20, 26–29]) that approximating the tail distribution of the queue contents by a geometric form may be quite accurate if the dominant singularity (i.e., the singularity with the smallest modulus) of $V(z)$ is a pole; as shown in [20], this will be the case if

$$\lim_{z \rightarrow \mathcal{R}} \frac{S(\lambda_1(z)^N)}{z} > 1, \quad (4.9)$$

where $\mathcal{R}$ is the minimum of the radii of convergence of the pgf's $S(G_{ij}(z))$, $1 \leq i, j \leq 2$. If each of the pgf's $S(\cdot)$ and $G_{ij}(\cdot)$ either is analytic in the entire complex plane or has a dominant singularity that is a pole, then $\mathcal{R}$ corresponds to a pole as well and the above inequality is always satisfied, since $S(\lambda_1(z)^N)$ will then tend to $+\infty$ for $z \rightarrow \mathcal{R}$. Only if $\mathcal{R}$ corresponds to a branch point, it may be so that this inequality is not satisfied; this calls for a case-by-case investigation of the corresponding tail behaviour of the queue contents and falls beyond the scope of the current paper.

Approximating the distribution of the queue contents by a geometric form corresponds to approximating $V(z)$ by

$$V(z) \cong \frac{C_v}{z - z_{0,v}} = -\frac{C_v}{z_{0,v}} \sum_{n=0}^{\infty} \left(\frac{z}{z_{0,v}} \right)^n, \quad (4.10)$$

where we are particularly interested in sufficiently large values of n . In this expression, $z_{0,v}$ is the pole with the smallest modulus of $V(z)$, which is a zero of the denominator for $m = N$ in expression (4.1) for $V(z)$, which yields

$$z_{0,v} = S(A_N(z_{0,v})) = S(\lambda_1(z_{0,v})^N), \quad (4.11)$$

and $z_{0,v}$ is a real and positive number larger than 1. We assume that this is a pole with multiplicity 1; note that the above formula can be readily extended to the case where $z_{0,v}$ has multiplicity larger than 1 should such a scenario occur. Furthermore, the constant C_v can be calculated from the residue theorem, leading to

$$\begin{aligned} C_v &= \lim_{z \rightarrow z_{0,v}} (z - z_{0,v})V(z) \\ &= (1 - \rho) \frac{z_{0,v}(z_{0,v} - 1)}{1 - N\lambda_1'(z_{0,v})\lambda_1(z_{0,v})^{N-1}S'(\lambda_1(z_{0,v})^N)} \xi_N(z_{0,v}). \end{aligned} \quad (4.12)$$

Combining these relations, we find the following approximate expression for the probability that the queue contents exceed a certain threshold n :

$$\Pr[v > n] \cong \frac{(1 - \rho)\xi_N(z_{0,v})}{1 - N\lambda_1'(z_{0,v})\lambda_1(z_{0,v})^{N-1}S'(\lambda_1(z_{0,v})^N)} z_{0,v}^{-n}. \quad (4.13)$$

Again, we can either use the exact value for $\xi_N(z_{0,v})$ calculated from (3.33) or the approximation calculated from (3.37).

5. The Unfinished Work

5.1. Derivation of the Probability Generating Function

The unfinished work in the queue at the beginning of a slot is defined as the amount of work in the queue at the beginning of the slot, which is the number of slots required to empty the queue if no more customers were to arrive during the subsequent slots. The amount of work required to process a customer whose service has not started yet is given by its service time described by $S(z)$. On the other hand, at the beginning of a slot, the customer currently being served (if any) still requires h slots before it is entirely sent, where h equals the residual service time. The unfinished work u_k at the beginning of slot $k + 1$ is therefore given by

$$u_k = \begin{cases} 0, & \text{if } v_k = 0, h_k = 0, \\ h_k + \sum_{i=1}^{v_k-1} s_i & \text{if } v_k > 0, \end{cases} \quad (5.1)$$

where each of the s_i 's represents a customer service time, and is therefore described by the pgf $S(z)$. Consequently, the corresponding steady-state pgf $U(z)$ is given by

$$U(z) = S(z)^{-1} \{ P(S(z), z, 1) + (S(z) - 1)P(0, 0, 1) \}. \quad (5.2)$$

The observation that $P(0, 0, 1) = \Pr[v = 0] = 1 - \rho$ can be written as

$$P(0, 0, 1) = \sum_{l=0}^N p_l = \sum_{m=0}^N \sum_{l=0}^N F_{lm}(S(z)) p_l, \quad (5.3)$$

together with the previous relations, leads to the following expression for the steady-state probability generating function describing the unfinished work at the beginning of an arbitrary slot:

$$U(z) = (1 - \rho) \sum_{m=0}^N \frac{(z - 1)A_m(S(z))}{z - A_m(S(z))} \xi_m(S(z)). \quad (5.4)$$

Again, due to $A'(1)S'(1) = NSp = \rho$ and $\xi_N(1) = \delta(N - m)$, note that $U(1) = 1$, as expected. We would like to point out that each of the denominators in the above expression also has a zero inside the unit disk $|z| \leq 1$, which leads to an additional linear equation for the boundary probabilities p_l (since $U(z)$ is bounded inside the complex unit circle $|z| < 1$), as explained in Section 3.4. However, note that if $\tilde{z}_{\{m\}}$ is a solution of $\tilde{z}_{\{m\}} = A_m(S(\tilde{z}_{\{m\}}))$ with $|\tilde{z}_{\{m\}}| \leq 1$, we then have that $S(\tilde{z}_{\{m\}}) = S(A_m(S(\tilde{z}_{\{m\}})))$, implying that

$$\begin{aligned} z_{\{m\}} &= S(\tilde{z}_{\{m\}}), \\ \tilde{z}_{\{m\}} &= A_m(z_{\{m\}}), \end{aligned} \quad (5.5)$$

where $z_{\{m\}}$ satisfies (3.35), and we obtain exactly the same set of linear equations as in (3.37) for the boundary probabilities.

5.2. Moments of the Unfinished Work

In a similar way to that in Section 4.2, if we define $\chi_m(z) \triangleq A_m(S(z))$ and take the appropriate derivatives with respect to z of expression (5.4) for $U(z)$, we find an expression for the mean and variance of the unfinished work:

$$\begin{aligned} \mathbf{E}[u] &= \chi'_N(1) + \frac{\chi''_N(1)}{2(1 - \chi'_N(1))} + L_p \xi'_N(1), \\ \mathbf{Var}[u] &= \chi''_N(1) + \chi'_N(1) - \chi'_N(1)^2 + \frac{\chi'''_N(1)}{3(1 - \chi'_N(1))} + \frac{\chi''_N(1)}{2(1 - \chi'_N(1))} \left(1 + \frac{\chi''_N(1)}{2(1 - \chi'_N(1))} \right) \\ &\quad + L_p^2 \xi''_N(1) + S''(1) \xi'_N(1) + L_p \xi'_N(1) - L_p^2 \xi'_N(1)^2 + 2L_p T'(1), \end{aligned} \quad (5.6)$$

where

$$\begin{aligned}
 \chi'_N(1) &= NpL_p = \rho, \\
 \chi''_N(1) &= \rho^2 \left(1 - \frac{1}{N}\right) + \rho \left(\frac{\lambda''_1(1)L_p}{p} + \frac{S''(1)}{L_p} \right), \\
 \chi'''_N(1) &= \rho^3 \left(1 - \frac{1}{N}\right) \left(1 - \frac{2}{N}\right) + 3\rho^2 \left(1 - \frac{1}{N}\right) \left(\frac{\lambda''_1(1)L_p}{p} + \frac{S''(1)}{L_p} \right) \\
 &\quad + \rho \left(\frac{\lambda'''_1(1)L_p^2}{p} + 3 \frac{\lambda''_1(1)S''(1)}{p} + \frac{S'''(1)}{L_p} \right).
 \end{aligned} \tag{5.7}$$

The derivatives of $\xi_N(z)$ and $T'(1)$ in the previous expressions are given by (4.4) (or (4.5) in case the approximations for the boundary probabilities are applied) and (4.8), respectively.

5.3. Tail Distribution of the Unfinished Work

Under similar assumptions to those outlined in Section 4.3, we can adopt a geometric tail approximation for the distribution of the unfinished work. We then obtain an approximation for the probability that the unfinished work exceeds a threshold n :

$$\Pr[u > n] \cong \frac{(1 - \rho)\xi_N(S(z_{0,u}))}{1 - NS'(z_{0,u})\lambda'_1(S(z_{0,u}))\lambda_1(S(z_{0,u}))^{N-1}} z_{0,u}^{-n}, \tag{5.8}$$

where the smallest pole $z_{0,u}$ of $U(z)$ is the solution of $z_{0,u} = \lambda_1(S(z_{0,u}))^N$, which is a real and positive quantity larger than 1, and, similar to (5.5), can also be calculated from $z_{0,u} = \lambda_1(z_{0,v})^N$. Also note that $S(z_{0,u}) = z_{0,v}$, implying that the above formula can be expressed in terms of the smallest pole $z_{0,v}$ of $V(z)$ as well.

6. The Customer Delay

6.1. Derivation of the Probability Generating Function

We define the customer delay d as the number of slots an arbitrary customer remains in the queue, including its service time, that is, the number of slots between the end of its arrival and departure slot, as depicted in Figure 1. Consider an arbitrary tagged customer that enters the queue during a slot, say slot I . Note that this slot is not an arbitrary slot but a slot where an arbitrary customer arrives (which for instance implies that at least one customer arrives). Due to the FCFS service discipline, the delay of a customer is determined by the amount of work in the queue upon the tagged customer's arrival. In particular, if we denote by $\hat{u}_I$ the unfinished work at the beginning of slot I and by f_I the number of customers that have

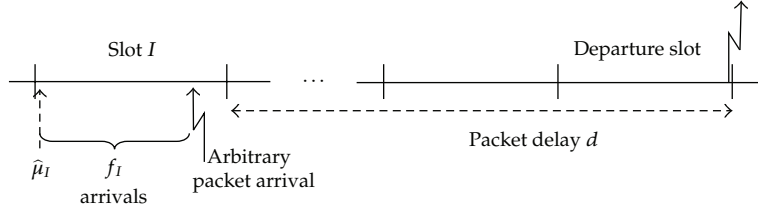


Figure 1: The delay of an arbitrary customer, with related quantities.

arrived during the same slot as the tagged one and will be served before it, then we find that the delay of the tagged customer is given by

$$d = (\hat{u}_I - 1)^+ + \sum_{i=1}^{f_I+1} s_i, \quad (6.1)$$

where the random variables $\{s_i, 1 \leq i \leq f_I + 1\}$ represent the service times of the tagged customer and the f_I customers that have arrived during its arrival slot and will be served before it, which are all described by the pgf $S(z)$. Let us denote by $G(z, x)$ the steady-state joint generating function of the random variables (u, e) describing the unfinished work at the beginning and the number of customer arrivals during an *arbitrary* slot k , that is,

$$G(z, x) = \mathbf{E}[z^u x^e], \quad (6.2)$$

and by $H(z, x)$ the steady state joint pgf of $(\hat{u}_I, f_k)$. Then, similar to [30], we can establish the following relation:

$$H(z, x) = \frac{G(z, 1) - G(z, x)}{Np(1 - x)}. \quad (6.3)$$

System equation (6.1) can be transformed into a relation between z -transforms, which together with (6.3), leads to an expression for the steady-state pgf $D(z)$ describing the delay of an arbitrary customer

$$\begin{aligned} D(z) &= \frac{S(z)}{z} \{H(z, S(z)) + (z - 1)H(0, S(z))\} \\ &= \frac{S(z)}{Npz} \left\{ \frac{G(z, 1) - G(z, S(z))}{1 - S(z)} + (z - 1) \frac{G(0, 1) - G(0, S(z))}{1 - S(z)} \right\}. \end{aligned} \quad (6.4)$$

On the other hand, since the unfinished work satisfies the system equation

$$u_k = (u_{k-1} - 1)^+ + \sum_{i=1}^{e_k} s_i \quad (6.5)$$

and assuming that a stochastic equilibrium is reached, this implies that the pgf $U(z)$ (given by (5.4)) can also be written as (again, all s_i 's are described by $S(z)$)

$$U(z) = z^{-1} \{G(z, S(z)) + (z-1)G(0, S(z))\}. \quad (6.6)$$

In combination with the previous expression for $D(z)$ (and making use of $U(z) = G(z, 1)$), this finally leads to

$$D(z) = \frac{S(z)}{z} \frac{S'(1)(z-1)}{S(z)-1} U_0(z), \quad (6.7)$$

where in view of $U(0) = \Pr[u=0] = \Pr[v=0] = 1-\rho$

$$U_0(z) = \mathbf{E}[z^u \mid u > 0] = \frac{(U(z) - U(0))}{\rho}. \quad (6.8)$$

This formula for $D(z)$, which expresses the delay of an arbitrary customer in terms of the unfinished work at the beginning of an arbitrary slot and which apparently is independent of the details of the customer arrival process (correlation, etc.) was also derived in [31] via an alternative analysis. From this result, it is not difficult to express the performance measures concerning the customer delay in terms of the corresponding performance measures for the unfinished work, as shown next.

6.2. Moments of the Customer Delay

By taking the first two order derivatives of (6.7)-(6.8) with respect to z for $z = 1$, we can establish the following relations between the mean and variance of the customer delay and the mean and variance of the unfinished work:

$$\begin{aligned} \mathbf{E}[d] &= L_p - 1 - \frac{S''(1)}{2L_p} + \frac{\mathbf{E}[u]}{\rho}, \\ \mathbf{Var}[d] &= \mathbf{Var}[s] - \frac{S'''(1)}{3L_p} - \frac{S''(1)}{2L_p} \left(1 - \frac{S''(1)}{2L_p}\right) + \frac{\mathbf{Var}[u]}{\rho} - (1-\rho) \left(\frac{\mathbf{E}[u]}{\rho}\right)^2. \end{aligned} \quad (6.9)$$

Through some simple calculations, one can verify that Little's theorem indeed holds, that is, $\mathbf{E}[d] = \mathbf{E}[v]/(Np)$.

6.3. Tail Distribution of the Customer Delay

Again, applying an analogous technique as the one used for determining the tail distribution of the queue contents and unfinished work, the asymptotic behavior of the customer delay tail distribution can be calculated from

$$D(z) \cong \frac{C_d}{z - z_{0,d}} = -\frac{C_d}{z_{0,d}} \sum_{n=0}^{\infty} \left(\frac{z}{z_{0,d}}\right)^n, \quad C_d \triangleq \lim_{z \rightarrow z_{0,d}} (z - z_{0,d})D(z), \quad (6.10)$$

where $z_{0,d}$ is the smallest pole of $D(z)$ which, in view of (6.7)-(6.8), leads to the observation that $z_{0,d} = z_{0,u}$, and the probability that the customer delay exceeds a threshold n is approximated by

$$\Pr[d > n] \cong \frac{1-\rho}{\rho} L_p \frac{(z_{0,u} - 1)}{S(z_{0,u}) - 1} \frac{S(z_{0,u}) \xi_N(S(z_{0,u}))}{1 - N \lambda'_1(S(z_{0,u})) S'(z_{0,u}) \lambda_1(S(z_{0,u}))} z_{0,u}^{-n-1}. \quad (6.11)$$

7. Some Numerical Examples

We will illustrate the results derived throughout the previous sections by a few numerical examples. Let us therefore consider a two-state ON/OFF MMBP arrival process for each source, which means that

$$G_{j1}(z) = 1 - \gamma + \gamma z, \quad G_{j2}(z) = 1, \quad 1 \leq j \leq 2, \quad (7.1)$$

that is, customers are generated at a rate γ during an ON period (corresponding with state S_1 in the modulating Markov chain) and no customers are generated during an OFF period (corresponding with state S_2). Let us also denote by L_1 and L_2 the mean length of an ON and OFF period, respectively, that is, $L_1 = 1/p_{12}$ and $L_2 = 1/p_{21}$. Furthermore, we will assume that customers have a constant service time equal to L_p slots, implying that $S(z) = z^{L_p}$. The load of the system is then given by

$$\rho = N L_p \gamma \frac{L_1}{L_1 + L_2}, \quad (7.2)$$

and any subset containing five parameters of the set $(N, \rho, \gamma, L_1, L_2, L_p)$ fully describes the customer arrival process in the queue. Such a model has for instance been used to effectively model a voice-over-IP (VoIP) packet arrival stream in a network node [3].

In the remainder we consider a multiplexer queue fed by $N = 16$ identical sources and focus on the customer delay behaviour, which is the primary performance measure for VoIP applications. Similar results can be readily obtained for the queue contents and unfinished work. For the arrival model described above, we have plotted the mean and variance of the the moments of the customer waiting time w (which is defined as the number of slots a customer waits in the queue before its service commences and is therefore equal to the customer delay minus the customer service time) versus ρ in case of $L_1 = L_2 = 8$ and for increasing customer service time $L_p = 1, 2, 4, 8, 16$ in Figures 2 and 4, respectively. Varying the value of the load in this case means that the arrival rate γ is adjusted. From these figures we can conclude that the moments of the customer waiting time strongly depend on the customer service time and increase for increasing values of L_p . In these figures, we have also included the curves when approximation (3.35) is used for the boundary probabilities (dashed lines), which leads to approximation (4.5) for the derivatives of $\xi_N(z)$; the absolute error between the exact and approximate results is so small that the latter are hardly visible in Figures 2 and 4. Therefore, in Figures 3 and 5, we have also plotted the relative error (RE) that is made when using the approximate formula for the mean and variance of the waiting time. Apparently, except for $L_p = 1$, the RE of mean and variance appears to be a decreasing function of ρ that approaches zero for $\rho = 1$. In addition,

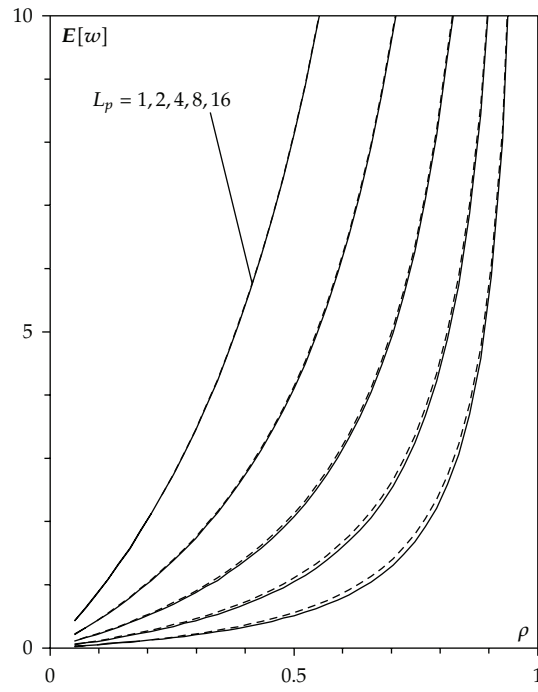


Figure 2: $E[w]$ versus ρ , $N = 16$, $L_1 = L_2 = 8$.

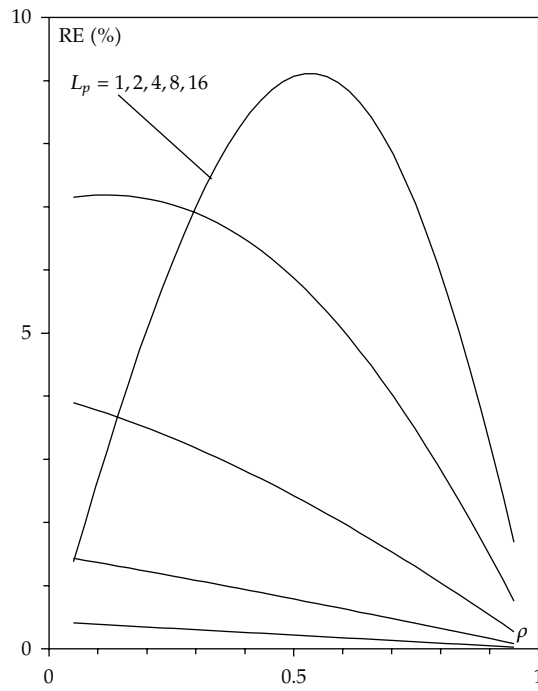


Figure 3: RE of $E[w]$ versus ρ , $N = 16$, $L_1 = L_2 = 8$.

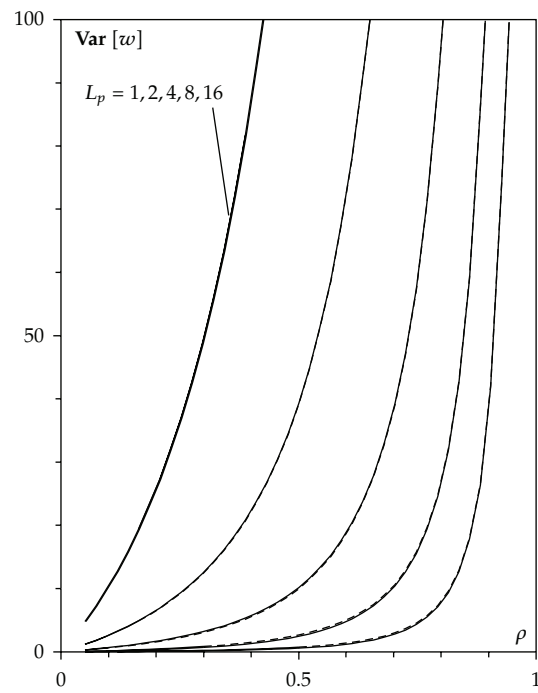


Figure 4: $\text{Var}[w]$ versus ρ , $N = 16$, $L_1 = L_2 = 8$.

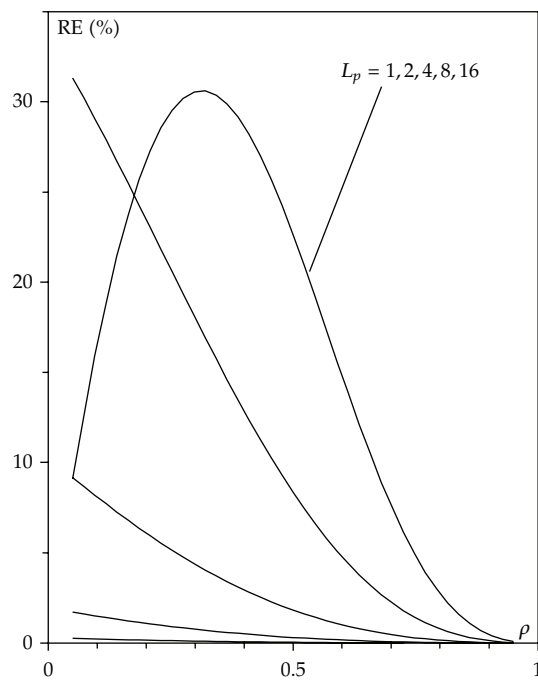


Figure 5: RE of $\text{Var}[w]$ versus ρ , $N = 16$, $L_1 = L_2 = 8$.

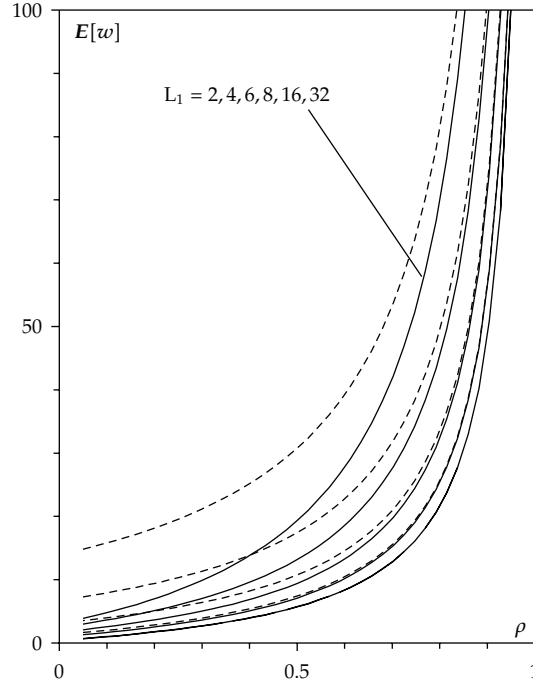


Figure 6: $E[w]$ versus ρ , $N = 16$, $\gamma = 0.05$, $L_p = 10$.

the larger the customer service times are, the smaller the relative error generally tends to be. We also observed that, while for the mean value the approximation always forms an upper bound compared to the exact value, this is not the case as far as the variance is concerned.

In Figure 6, we consider constant values for $\gamma = 0.05$ and $L_p = 10$ and increasing values of the mean length L_1 of an active period. Hence, the probability $L_1/(L_1 + L_2)$ that a source is in the ON state during an arbitrary slot tends to 0 for $\rho \rightarrow 0$. Clearly, the mean value of the customer waiting time, which is plotted versus ρ in these figures, strongly depends on the mean length of an active period. In addition, we observe from these curves that the error that is made by using the approximation for the boundary probabilities (dashed line) can be significant when the load is low and the arrival rate γ and L_1 are high. Another feature is that the mean waiting time not necessarily tends to 0 for $\rho \rightarrow 0$. Indeed, even under these conditions, we sporadically encounter the situation where a source becomes active and generates customers at a rate γ during a geometrically distributed ON period with mean length L_1 , and the intersection with the ordinate $\rho = 0$ designates the mean waiting time for customers generated by such a single source.

In Figures 7, 8 and 9, we have plotted the tail distribution of the customer waiting time for a wide range of system parameters. Each curve for the exact value of the parameters of the geometric tail behaviour (full line) is accompanied by the corresponding approximation (dashed line) obtained by using approximation (3.37) for $\xi_N(z)$. We may conclude that, for the wide range of system parameters considered in these figures, the approximate results are extremely accurate. Furthermore, it is observed that the approximation forms an upper bound, and, consequently, is on the safe side. The only numerical calculation required for this

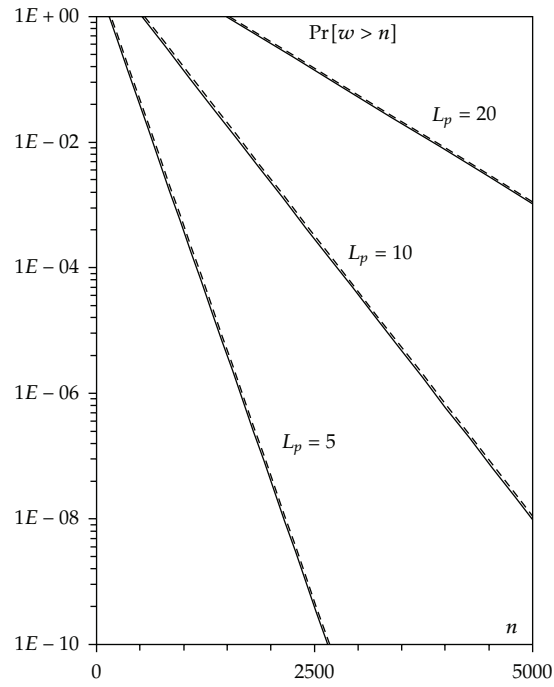


Figure 7: $\Pr[w > n]$ versus n , $N = 16$, $\rho = 0.8$, $\gamma = 0.1$, $L_1 = 50$.

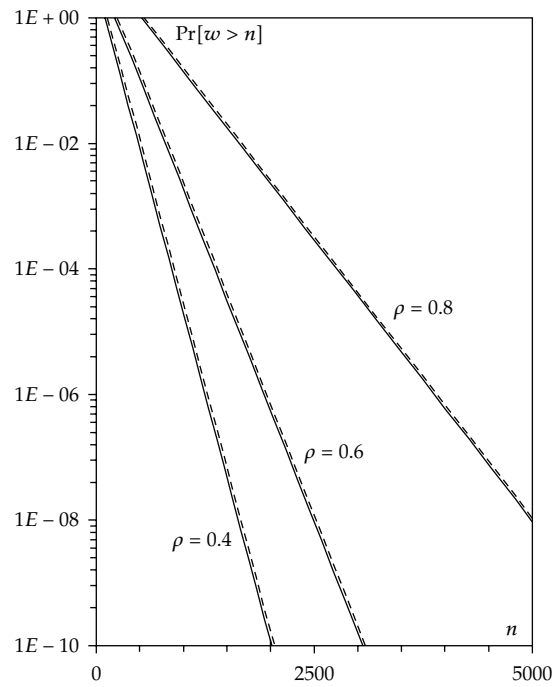


Figure 8: $\Pr[w > n]$ versus n , $N = 16$, $\gamma = 0.1$, $L_p = 10$, $L_1 = 50$.

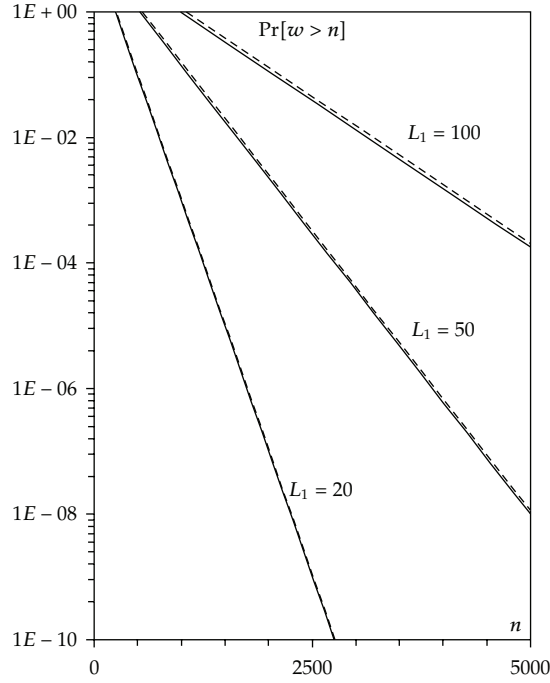


Figure 9: $\Pr[w > n]$ versus n , $N = 16$, $\rho = 0.8$, $\gamma = 0.1$, $L_p = 10$.

upper bound is that of the dominant pole $z_{0,u}$ of the pgf describing the steady-state waiting time distribution. Insertion of the system parameters in the corresponding formulas then immediately yields these results.

In Figure 7, we have plotted the probability that the customer waiting time exceeds a certain threshold n versus n , for a multiplexer queue with $N = 16$ sources, a system load $\rho = 0.8$, an arrival rate $\gamma = 0.1$ during an ON period, and a mean length of an ON period $L_1 = 50$, for varying values of the customer service time L_p . Again, as was also observed when discussing the moments, for these parameter values the results for the quantile of the waiting time (defined as the smallest value of $n(x)$ for which $\Pr[w > n(x)] < 10^{-x}$) significantly depend on the customer service time. From this figure, it also becomes apparent that the quantile of the waiting time is roughly proportional to L_p , and a wide range of case studies has revealed that this observation is not limited to the case of constant customer service times.

In Figure 8, we show the results for the customer waiting time in case of $N = 16$, $\gamma = 0.1$, $L_1 = 50$ and $L_p = 10$ and increasing values of the load (which means that the arrival rate γ during the ON periods increases), while the load is fixed to $\rho = 0.8$ and $L_1 = 20, 50, 100$ in Figure 9. As can be expected, the slope of the customer waiting time tail distribution increases as both ρ and L_1 increase. We also should point out that, whereas the accuracy of the approximate results for the moments of the customer waiting time—calculated by invoking the approximation for the boundary probabilities—deteriorates for increasing values of L_1 (and L_2) (e.g., Figure 6), this is not so as far as the tail distribution of the customer waiting time is concerned, and the approximate data that we thus compute remain highly accurate.

8. Conclusion

In this paper we have investigated and analysed the discrete-time (2-state) D-BMAP/G/1 queueing model. By combining a solution method that was previously developed for the D-BMAP/D/c queueing system, with the supplementary variable technique that allows us to cope with generally distributed customer service times, we were able to derive an expression for the steady-state pgfs of the queue contents, unfinished work, and customer delay. This has led to (semi)closed-form expressions for a number of interesting performance measures such as the mean value, variance, and tail distribution of these quantities. In addition to the analysis of the system, a lot of emphasis was put on finding closed-form expressions for these quantities that reduce all numerical calculations to an absolute minimum, at the expense of introducing approximations for a number of boundary probabilities that must normally be calculated by solving a—potentially large—set of linear equations. The only numerical computation that needs to be carried out is that of a dominant pole, situated on the positive real axis, when one calculates tail probabilities, and the complexity does not depend on the size of the state space of the model under consideration. To conclude, the accuracy of this approach, as well as the impact of the system parameters on the customer delay performance, was evaluated through a considerable set of numerical examples.

Appendix

The Eigenvalues and Eigenvectors of $Q(z)$

As it became clear from the analysis throughout this report, an important role is played by the eigenvalues and eigenvectors of $Q(z)$. Let us denote by $\Lambda(z)$ the 2×2 diagonal matrix containing the eigenvalues $\lambda_i(z)$ and by $U(z)$ and $W(z)$ the 2×2 matrices containing the corresponding right column and left row eigenvectors, respectively, that is,

$$W(z)Q(z)U(z) = \Lambda(z), \quad U(z)W(z) = W(z)U(z) = I, \quad (A.1)$$

with I the 2×2 diagonal unity matrix. The normalization constants of the left row and right column eigenvectors are chosen such that

$$U(z)\mathbf{1} = W(z)\mathbf{1} = \mathbf{1}. \quad (A.2)$$

We will denote by $\mathbf{u}_j(z)$ ($\mathbf{w}_j(z)$) the right column (left row) eigenvector corresponding to $\lambda_j(z)$, $1 \leq j \leq 2$.

Taking into account expression (2.1) for $Q(z)$, the characteristic equation of this matrix can be written as

$$\lambda(z)^2 - (q_{11}(z) + q_{22}(z))\lambda(z) + q_{11}(z)q_{22}(z) - q_{12}(z)q_{21}(z) = 0. \quad (A.3)$$

Solution of the characteristic equation of $\mathbf{Q}(z)$ leads to

$$\frac{\lambda_1(z)}{\lambda_2(z)} = \frac{q_{11}(z) + q_{22}(z) \pm \left((q_{11}(z) - q_{22}(z))^2 + 4q_{12}(z)q_{21}(z) \right)^{1/2}}{2}. \quad (\text{A.4})$$

In addition, solving (2.16) and (3.1) yields the following expression for $\mathbf{U}(z)$:

$$\mathbf{U}(z) = \begin{bmatrix} \frac{(\lambda_1(z) - q_{11}(z) + q_{21}(z))q_{12}(z)}{(q_{11}(z) - \lambda_1(z))(\lambda_2(z) - \lambda_1(z))} & \frac{(\lambda_2(z) - q_{11}(z) + q_{21}(z))q_{12}(z)}{(q_{11}(z) - \lambda_2(z))(\lambda_1(z) - \lambda_2(z))} \\ \frac{(\lambda_1(z) - q_{11}(z) + q_{21}(z))}{(\lambda_1(z) - \lambda_2(z))} & \frac{(\lambda_2(z) - q_{11}(z) + q_{21}(z))}{(\lambda_2(z) - \lambda_1(z))} \end{bmatrix}. \quad (\text{A.5})$$

In a similar way, we can also calculate the elements of $\mathbf{W}(z)$; however, the explicit expression of this matrix is not required in our derivations. Note that

$$\begin{aligned} \lambda_1(1) &= 1, \\ \lambda_2(1) &= 1 - p_{12} - p_{21}. \end{aligned} \quad (\text{A.6})$$

$\lambda_1(z)$ is the Perron-Fröbenius (P-F) eigenvalue of $\mathbf{Q}(z)$ and is an important quantity for the analysis in this paper. Similarly, $\mathbf{u}_1(z)$ will be referred to as the P-F eigenvector of $\mathbf{Q}(z)$.

In our expressions for the mean and variance of the queue contents, unfinished work and customer delay, we also require the first-, second-, and third-order derivatives with respect to z for $z = 1$ of the P-F eigenvalue. Taking the appropriate derivatives of the characteristic equation (A.3), we obtain with (A.6)

$$\begin{aligned} \lambda'_1(1) &= \sum_{i=1}^2 \sigma_i \sum_{j=1}^2 q'_{ij}(1) \hat{= } p, \\ \lambda''_1(1) &= \sum_{i=1}^2 \sigma_i \sum_{j=1}^2 q''_{ij}(1) + 2 \frac{\lambda'_1(1)\lambda'_2(1) + q'_{12}(1)q'_{21}(1) - q'_{11}(1)q'_{22}(1)}{p_{12} + p_{21}}, \\ \lambda'''_1(1) &= \sum_{i=1}^2 \sigma_i \sum_{j=1}^2 q'''_{ij}(1) + 3 \frac{\lambda''_1(1)\lambda'_2(1) + \lambda'_1(1)\lambda''_2(1) + q''_{12}(1)q'_{21}(1) + q'_{12}(1)q''_{21}(1)}{p_{12} + p_{21}} \\ &\quad - 3 \frac{q''_{11}(1)q'_{22}(1) + q'_{11}(1)q''_{22}(1)}{p_{12} + p_{21}}, \end{aligned} \quad (\text{A.7})$$

where p was defined in (2.16). The n th derivative of $\lambda_2(z)$ with respect to z for $z = 1$ is readily calculated in terms of the n th derivative of $\lambda_1(z)$ from the relation

$$\lambda_1(z) + \lambda_2(z) = q_{11}(z) + q_{22}(z). \quad (\text{A.8})$$

Similarly, we also require the first- and second-order derivatives with respect to z for $z = 1$ of the P-F eigenvector. First of all, we point out that, due to (A.5), $\mathbf{U}(1)$ satisfies

$$\mathbf{U}(1) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}. \quad (\text{A.9})$$

Then, (Strictly speaking, the second column of $\mathbf{U}(z)$ is no longer an eigenvector of $\mathbf{Q}(z)$ for $z = 1$. However, it is not difficult to verify that all the results derived throughout this paper still hold when taking the limit $z = 1$.) taking the appropriate derivatives of the equations generated by (A.2) and the relation $\mathbf{Q}(z)\mathbf{U}(z) = \mathbf{U}(z)\mathbf{\Lambda}(z)$, we obtain

$$\begin{aligned} \mathbf{U}'(1) &= \frac{\delta}{p_{12}} \begin{bmatrix} -\sigma_2 & \sigma_2 \\ \sigma_1 & -\sigma_1 \end{bmatrix}, \\ \mathbf{U}''(1) &= \frac{\gamma_1}{p_{12}} \begin{bmatrix} -\sigma_2 & \sigma_2 \\ \sigma_1 & -\sigma_1 \end{bmatrix} + \frac{\gamma_2}{p_{12}} \begin{bmatrix} -\sigma_2 & \sigma_2 \\ -\sigma_2 & \sigma_2 \end{bmatrix}, \end{aligned} \quad (\text{A.10})$$

where

$$\begin{aligned} \delta &= p - (q'_{11}(1) + q'_{12}(1)), \\ \gamma_1 &= \lambda''_1(1) - 2 \frac{\delta p}{p_{12} + p_{21}} - (q''_{11}(1) + q''_{12}(1)) - 2(q'_{11}(1)u'_{11}(1) + q'_{12}(1)u'_{21}(1)), \\ \gamma_2 &= -2 \frac{\delta \lambda'_2(1)}{p_{12} + p_{21}} + 2(q'_{11}(1)u'_{11}(1) + q'_{12}(1)u'_{21}(1)). \end{aligned} \quad (\text{A.11})$$

Disclosure

J. Walraevens and D. Fiems are Postdoctoral Fellows of the Research Foundation-Flanders (F.W.O.).

References

- [1] Li and Jon W. Mark, "Traffic characterization for integrated services networks," *IEEE Transactions on Communications*, vol. 38, no. 8, pp. 1231–1243, 1990.
- [2] J.-A. Zhao, B. Li, and I. Ahmad, "Traffic modeling for layered video," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '03)*, pp. 497–500, Baltimore, Md, USA, 2003.
- [3] D. de Vleeschauwer, A. van Moffaert, M. Buchli et al., "Determining the tolerable load generated by a set of customer-based phones on a multiplexing node," in *Proceedings of the 17th International Teletraffic Congress (ITC '01)*, pp. 433–444, Salvador da Bahia, Brazil, 2001.
- [4] F. Guillemin, J. Boyer, and A. Dupuis, "Burstiness in broadband integrated networks," *Performance Evaluation. An International Journal*, vol. 15, no. 3, pp. 163–176, 1992.
- [5] M. R. Izquierdo and D. S. Reeves, "Survey of statistical source models for variable-bit-rate compressed video," *Multimedia Systems*, vol. 7, no. 3, pp. 199–213, 1999.
- [6] N. Ansari, H. Liu, Y. Q. Shi, and H. Zhao, "On modeling MPEG video traffics," *IEEE Transactions on Broadcasting*, vol. 48, no. 4, pp. 337–347, 2002.

- [7] A. Lombardo, G. Morabito, and G. Schembra, "Modeling intramedia and intermedia relationships in multimedia network analysis through multiple timescale statistics," *IEEE Transactions on Multimedia*, vol. 6, no. 1, pp. 142–157, 2004.
- [8] Hashida, Takahashi, and Shimogawa, "Switched batch Bernoulli process (SBBP) and the discrete-time SBBP/G/1 queue with application to statistical multiplexer performance," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 3, pp. 394–401, 1991.
- [9] N. C. Hock, B. Lizhou, and Z. Liren, "Queue length solutions for an ATM buffer with MMBP arrivals," *Computer Communications*, vol. 20, no. 10, pp. 878–883, 1997.
- [10] D. Park and H. G. Perros, "m-MMBP characterization of the departure process of an m-MMBP/Geo/1/K queue," in *Proceedings of the 14th ITC*, pp. 75–84, Juan-Les-Pins, France, 1994.
- [11] T. Tsuchiya and Y. Takahashi, "On discrete-time single-server queues with Markov modulated batch Bernoulli input and finite capacity," *Journal of the Operations Research Society of Japan*, vol. 36, no. 1, pp. 29–45, 1993.
- [12] C. Blondia and O. Casals, "Statistical multiplexing of VBR sources: a matrix-analytic approach," *Performance Evaluation*, vol. 16, no. 1-3, pp. 5–20, 1992.
- [13] I. Frigui, A. S. Alfa, and X. Xu, "Algorithms for computing waiting time distributions under different queue disciplines for the D-BMAP/PH/1," *Naval Research Logistics. A Journal Dedicated to Advances in Operations and Logistics Research*, vol. 44, no. 6, pp. 559–576, 1997.
- [14] C. Herrmann, "Complete analysis of the discrete time finite DBMAP/G/1/N queue," *Performance Evaluation*, vol. 43, no. 2-3, pp. 95–121, 2001.
- [15] K. Kang and C. Kim, "Performance analysis of statistical multiplexing of heterogeneous discrete-time Markovian arrival processes in an ATM network," *Computer Communications*, vol. 20, no. 11, pp. 970–978, 1997.
- [16] B. Kim and J. Kim, "Queue size distribution in a discrete-time D-BMAP/G/1 retrial queue," *Computers & Operations Research*, vol. 37, no. 7, pp. 1220–1227, 2010.
- [17] H. W. Lee, J. M. Moon, B. K. Kim, J. G. Park, and S. W. Lee, "A simple eigenvalue method for low-order D-BMAP/G/1 queues," *Applied Mathematical Modelling*, vol. 29, no. 3, pp. 277–288, 2005.
- [18] Li, "A general solution technique for discrete queueing analysis of multimedia traffic on ATM," *IEEE Transactions on Communications*, vol. 39, no. 7, pp. 1115–1132, 1991.
- [19] S. K. Samanta, U. C. Gupta, and R. K. Sharma, "Analyzing discrete-time D-BMAP/G/1/N queue with single and multiple vacations," *European Journal of Operational Research*, vol. 182, no. 1, pp. 321–339, 2007.
- [20] B. Steyaert, *Analysis of generic discrete-time queue models with irregular customer arrival patterns*, Ph.D. thesis, Gent, Belgium, 2008, <https://biblio.ugent.be/record/471285>.
- [21] B. Steyaert, Y. Xiong, and H. Bruneel, "An efficient solution technique for discrete-time queues FED by heterogeneous traffic," *International Journal of Communication Systems*, vol. 10, no. 2, pp. 73–86, 1997.
- [22] G. Latouche and V. Ramaswami, *Introduction to matrix analytic methods in stochastic modeling*, Cambridge University Press, Cambridge, UK, 2002.
- [23] M. F. Neuts, *Matrix-geometric solutions in stochastic models*, vol. 2 of *Johns Hopkins Series in the Mathematical Sciences*, Johns Hopkins University Press, Baltimore, Md., 1981, An algorithmic approach.
- [24] J. W. Baek, H. W. Lee, S. W. Lee, and S. Ahn, "A factorization property for BMAP/G/1 vacation queues under variable service speed," *Annals of Operations Research*, vol. 160, pp. 19–29, 2008.
- [25] H. Bruneel, "Performance of discrete-time queueing systems," *Computers and Operations Research*, vol. 20, no. 3, pp. 303–320, 1993.
- [26] B. Steyaert and H. Bruneel, "Analysis of ATM switching modules with channel grouping in a bursty-source environment," in *Proceedings of the IFIP TC6/ICCC International Conference on Integrated broadband Communication Networks and Services (IBCN&S '94)*, pp. 383–394, Copenhagen, Denmark, 1994.
- [27] G. U. Hwang and B. D. Choi, "Closed-form expressions on the geometric tail behavior of statistical multiplexers with heterogeneous traffic," *IEEE Transactions on Communications*, vol. 46, no. 12, pp. 1575–1579, 1998.
- [28] Sohraby, "On the theory of general ON-OFF sources with applications in high-speed networks," in *Proceedings of the IEEE Conference on Computer Communications, (INFOCOM '93)*, vol. 2, pp. 401–410, 1993.
- [29] Y. Xiong and H. Bruneel, "A simple approach to obtain tight upper bounds for the asymptotic queueing behavior of statistical multiplexers with heterogeneous traffic," *Performance Evaluation*, vol. 22, no. 2, pp. 159–173, 1995.

- [30] B. Steyaert, H. Bruneel, and Y. Xiong, "A general relationship between buffer occupancy and delay in discrete-time multiserver queueing models, applicable in ATM networks," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '93)*, vol. 3, pp. 1250–1258, San Francisco, Calif., USA, 1993.
- [31] B. Vinck and H. Bruneel, "General relationships between queue length and delay in discrete-time single-server queues," in *Proceedings of the COST 257 Management Committee Meeting*, Espoo, Finland, 1997.