Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 324604, 26 pages
doi:10.1155/2012/324604

Research Article

A Tandem BMAP/$G/1 \to \bullet/M/N/0$ Queue with Group Occupation of Servers at the Second Station

Chesoong Kim,[1] Alexander Dudin,[2] Valentina Klimenok,[2] and Olga Taramin[2]

[1] *Department of Industrial Engineering, Sangji University, Wonju, Kangwon 220-702, Republic of Korea*
[2] *Department of Applied Mathematics and Computer Science, Belarusian State University, 4 Nezavisimosti Avenue, 220030 Minsk, Belarus*

Correspondence should be addressed to Chesoong Kim, dowoo@sangji.ac.kr

Received 21 July 2011; Accepted 13 September 2011

Academic Editor: M. D. S. Aliyu

We consider a two-stage tandem queue with single-server first station and multiserver second station. Customers arrive to Station 1 according to a batch Markovian arrival process (BMAP). A batch may consist of heterogeneous customers. The type of a customer is determined upon completion of a service at Station 1. The customer's type is classified based on the number of servers required to process the request of the customer at Station 2. If the required number of servers is not available, the customer may leave the system forever or block Station 1 by waiting for the required number of servers. We determine the stationary distribution of the system states at embedded epochs and derive the Laplace-Stieltjes transform of the sojourn time distribution. Some key performance measures are calculated, and illustrative numerical results are presented.

1. Introduction

Queueing networks are widely used in capacity planning and performance evaluation of computer and communication systems, service centers, and manufacturing systems among several others. Some examples of their application to real systems can be found in [1]. Tandem queues can be used for modeling real-life two-node networks as well as for the validation of general decomposition algorithms in networks (see, e.g., [2, 3]). Thus, tandem queueing systems have found much interest in the literature. An extensive survey of early papers on tandem queues can be seen in [4]. Most of these papers are devoted to exponential queueing models. Over the last two decades or so, the efforts of many investigators in tandem queues were in weakening the distribution assumptions on the service times as well as on the arrivals. In particular, the arrival process should be able to capture any correlation and burstiness that are commonly seen in the traffic of modern communication networks [3]. Such an arrival process was introduced in [5] and ever since this process is referred to as a batch Markovian arrival process (BMAP). In this paper, we deal with a tandem queue under the assumption that the customers arrive according to a BMAP.

Tandem queues with the BMAP input were considered in [6–10]. The papers [6, 7, 10] are devoted to the $\mathrm{MAP}/\mathrm{PH}/1 \rightarrow \bullet/G/1$ system with blocking. In [8], the tandem queues $\mathrm{BMAP}/G/1/N \rightarrow \bullet/\mathrm{PH}/1/M-1$ with losses are studied. The tandem-queue of the $\mathrm{BMAP}/G/1 \rightarrow \bullet/\mathrm{PH}/1/M-1$ type with losses and feedback has been studied in [9].

In the present paper, we consider a $\mathrm{BMAP}/G/1 \rightarrow \bullet/M/N/0$ tandem queueing system where the (possibly heterogeneous) customers arrive in batches of random sizes to Station 1. Here the customers receive service individually, and upon completion of a service the customer's type is determined. This type identification is necessary to determine the nature of service, if any, offered at Station 2. The customer's type is classified based on the number of servers (resources) required to process the request of the customer. The simultaneous initiation or occupation of several servers to a customer's request is typical for the so-called nonelastic traffic in communication networks. If the required number of servers is not available at that instant of the request, the customer either leaves the system forever, or awaits until the requirement is met through the release of the sufficient number of servers. In the latter case, Station 1 will be blocked.

Possible applications of the tandem queue under study lie in the modeling of the distributed server application or web server application, see, for example, [11]. Station 1 is interpreted as an authentication or an access step while Station 2 represents the computing step or data base server if the processing of a job is produced by several parallel threads. This tandem queue can model also multiaddress transmission of information. Station 1 is interpreted as a transmission channel while Station 2 regulates the transmission rate by providing necessary transmission windows (timers that are switched on at the moment of a message transmission and switched off when the receipt of this message is acknowledged or time-out expires). The performance evaluation of wireless IP networks providing heterogeneous multimedia services with different QoS demands, (see [12, Chapter 8]), is the other possible application of the model under study.

In this paper, we derive the stability condition of the model under study, and briefly touch calculation of the stationary distribution of the system states at the service completion epochs at Station 1 and calculation of the system performance measures. Furthermore, we derive the Laplace-Stieltjes transform of the virtual and the actual sojourn time distributions at both stations and in the whole system. The procedures for calculation of the moments of the virtual sojourn time distribution and the mean actual sojourn time are discussed. Some numerical results illustrating the behavior of the system characteristics are presented. The problem of optimal design is numerically investigated.

To the best of our knowledge, the results of our paper are novel even for the case of homogeneous customers. The most important and valuable, from the mathematical point of view, result concerns the sojourn time distribution. Previously, the sojourn time distribution in tandem queues with MAP input was considered only in [13, 14]. There, the service time distribution at both the single-server stations is of phase type which allows the authors to model the sojourn time as the time until absorption in suitably defined quasi-birth-and-death processes and continuous-time Markov chains. Because we assume general service time distribution at Station 1, we need to analyze a more complicated stochastic process.

The rest of the paper is organized as follows. In Section 2, the mathematical model is described. In Section 3, the results concerning the stationary distribution of the embedded Markov chain in Station 1 service completion epochs are presented. In Section 4, we focus on the analysis of the virtual and actual sojourn time distributions and their moments. In Section 5, the numerical results are presented. The paper is concluded with Section 6. Appendices contain auxiliary results, proofs, and formulas useful for computations.

2. The Mathematical Model

We consider a tandem queue consisting of two stations, say, Station 1 and Station 2. We assume that there is no buffer between the two stations. Station 1 is represented by the BMAP/G/1 queue. That is, the arrivals to Station 1 are described by a BMAP. The BMAP is defined by the underlying process $\nu_t,\ t \geq 0$, which is an irreducible continuous time Markov chain with state space $\{0, \ldots, W\}$, and with the matrix generating function $D(z) = \sum_{k=0}^{\infty} D_k z^k,\ |z| \leq 1$. Arrivals occur only at epochs of the jumps in the underlying process $\nu_t,\ t \geq 0$. The intensities of the transitions of the process ν_t accompanied by a batch of size k are defined by the matrices $D_k,\ k \geq 0$. The matrix $D(1)$ is the infinitesimal generator of the process ν_t. The stationary distribution vector $\boldsymbol{\theta}$ of this process satisfies the equations $\boldsymbol{\theta} D(1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1$, where $\mathbf{e}$ is a column vector consisting of 1′s, and $\mathbf{0}$ is a row vector of 0′s.

The average intensity λ (fundamental rate) of the BMAP is given by $\lambda = \boldsymbol{\theta} D'(z)|_{z=1}\mathbf{e}$. We assume that $\lambda < \infty$. The average intensity λ_b of group arrivals is defined by $\lambda_b = \boldsymbol{\theta}(-D_0)\mathbf{e}$. The coefficient of variation, c_{var}, of intervals between successive group arrivals is defined by $c_{\mathrm{var}}^2 = 2\lambda_b \boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$. The coefficient of correlation c_{cor} of the successive intervals between group arrivals is given by $c_{\mathrm{cor}} = (\lambda_b \boldsymbol{\theta}(-D_0)^{-1}(D(1) - D_0)(-D_0)^{-1}\mathbf{e} - 1)/c_{\mathrm{var}}^2$. For more information about the BMAP and related research see, for example, [5, 15].

All arriving customers enter into Station 1. The successive service times of customers at Station 1 are independent random variables with general distribution $B(t)$, Laplace-Stieltjes transform $\beta(s) = \int_0^{\infty} e^{-st} dB(t)$, and finite first moment $b_1 = \int_0^{\infty} t\, dB(t)$.

After receiving a service at Station 1, the customer proceeds to Station 2. At this station, there are N identical servers. Each of these servers offers services that are exponentially distributed with parameter μ. Customers are heterogeneous with respect to the number of servers that are required to process a customer at Station 2. With probability $q_m, q_m \geq 0, m = \overline{0, N}, \sum_{m=0}^{N} q_m = 1$, the customer will require exactly m servers to provide a service at Station 2 and will be called *type m customer*. Here and in the sequel, notation such as $m = \overline{0, N}$, means that m assumes values from the set $\{0, 1, \ldots, N\}$. Note that customers who are all in the same batch (at the time of arriving) may belong to different types after receiving service at Station 1. Type 0 customer leaves the system for good after the service at Station 1. We assume that $q_0 \neq 1$. Otherwise, the queue under consideration will be reduced to the BMAP/G/1 queue which has been studied extensively.

If the customer is of type m, $m = \overline{1, N}$, and the required number of servers is available, the customer's service will begin immediately. Each of these m servers processes the customer's request independently of the others, and furthermore any server who becomes free after completing his/her share of the processing will be available to process waiting or future customers' requests.

If the required number of servers is not available, with probability $\gamma,\ 0 \leq \gamma \leq 1$, the customer will choose to leave the system for good and with probability $1 - \gamma$ will decide to wait until the required number of servers is available. In the latter case, the customer will block Station 1 since we assume that there is no buffer between the two stations. Such an assumption of blocking and loss will allow us to unify these two classes of models which are studied separately in the literature.

In the following, we are interested in the steady state analysis of the model under study. For further use in the sequel, we introduce the following notation:

(i) I is an identity matrix of appropriate dimension;

(ii) $\otimes$ and $\oplus$ are symbols of the Kronecker product and sum of matrices;

(iii) $\widetilde{D}_k = I_{N+1} \otimes D_k,\ k \geq 0,\ \widetilde{D}(z) = \sum_{k=0}^{\infty} \widetilde{D}_k z^k,\ |z| \leq 1$;

(iv) $P(j,t), j \geq 0$, is a matrix function defined by the expansion $\sum_{j=0}^{\infty} P(j,t) z^j = e^{D(z)t}$;

(v) $F(t) = (F_{r,r'}(t))_{r,r'=\overline{0,N}}$, where $F_{r,r'}(t) = 0$ for $r \leq r'$ and for $r > r'$, $F_{r,r'}(t)$ is the generalized Erlang distribution function with the Laplace-Stieltjes transform $f_{r,r'}(s) = \prod_{l=r'+1}^{r} l\mu(l\mu + s)^{-1}$;

(vi) $Q_m,\ m = \overline{1,4}$, are square matrices: $Q_2 = \operatorname{diag}\{\sum_{m=N-r+1}^{N} q_m, r = \overline{0,N}\}$,

$$Q_1 = \begin{pmatrix} q_0 & q_1 & \cdots & q_N \\ 0 & q_0 & \cdots & q_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_0 \end{pmatrix}, \qquad Q_3 = \begin{pmatrix} 0 & \cdots & 0 & q_N \\ 0 & \cdots & 0 & q_{N-1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & q_0 \end{pmatrix}, \qquad Q_4 = \begin{pmatrix} q_0 & q_1 & \cdots & q_{N-1} & q_N \\ q_0 & q_1 & \cdots & q_{N-1} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ q_0 & 0 & \cdots & 0 & 0 \end{pmatrix}; \tag{2.1}$$

(vii) $\widetilde{Q}_m = Q_m \otimes I_{\overline{W}},\ m = \overline{1,3},\ \overline{W} = W + 1$;

(viii) $\widehat{Q} = \widetilde{Q}_1 + \gamma \widetilde{Q}_2 + (1-\gamma)\int_0^{\infty} (dF(t) \otimes e^{D_0 t}) \widetilde{Q}_3,\ Q = Q_1 + \gamma Q_2 + (1-\gamma) E Q_3$;

(ix)

$$E = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}, \quad \widehat{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \quad \widetilde{I} = I - \widehat{I},\ \hat{\mathbf{e}} = (1, 0, \ldots, 0). \tag{2.2}$$

3. The Stationary Distribution of the Embedded Markov Chain

Let t_n denote the time of the nth service completion at Station 1. Consider the process $\xi_n = \{i_n,\ r_n,\ \nu_n\}$, $n \geq 1$, where $i_n, i_n \geq 0$, is the number of customers at Station 1 (not counting the blocked customer, if any) at epoch $t_n + 0$; $r_n, r_n = \overline{0,N}$, is the number of busy servers at Station 2 at epoch $t_n - 0$; $\nu_n, \nu_n = \overline{0,W}$, is the state of the BMAP at epoch t_n.

It is easy to verify that the process $\xi_n = \{i_n,\ r_n,\ \nu_n\}$, $n \geq 1$, is a Markov chain. Enumerating the states of this Markov chain in lexicographic order, and denoting by $P_{l,k},\ l,k \geq 0$, the square matrix of order $(W+1)(N+1)$ governing the transition probabilities of the chain from the set of states $\{l,\cdot,\cdot\}$ to the set $\{k,\cdot,\cdot\}$, the following lemma gives the entries of the transition probability matrix of the Markov chain ξ_n.

Lemma 3.1. *The transition probability matrix of the chain ξ_n, $n \geq 1$, has the following block structure:*

$$P = (P_{l,k})_{l,k\geq 0} = \begin{pmatrix} C_0 & C_1 & C_2 & C_3 & \cdots \\ Y_0 & Y_1 & Y_2 & Y_3 & \cdots \\ 0 & Y_0 & Y_1 & Y_2 & \cdots \\ 0 & 0 & Y_0 & Y_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{3.1}$$

where

$$C_i = \sum_{k=1}^{i+1} \left[-\widehat{Q}(\Delta \oplus D_0)^{-1} \widetilde{D}_k + (1-\gamma) F_k \widetilde{Q}_3 \right] \Omega_{i-k+1},$$

$$Y_i = \left(\widetilde{Q}_1 + \gamma \widetilde{Q}_2 \right) \Omega_i + (1-\gamma) \sum_{k=0}^{i} F_k \widetilde{Q}_3 \Omega_{i-k},$$

$$\Omega_j = \int_0^\infty e^{\Delta t} \otimes P(j,t)\, dB(t), \qquad F_j = \int_0^\infty dF(t) \otimes P(j,t), \quad j \geq 0, \tag{3.2}$$

$$\Delta = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ \mu & -\mu & 0 & \cdots & 0 & 0 \\ 0 & 2\mu & -2\mu & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & N\mu & -N\mu \end{pmatrix}.$$

Proof. First, we write the transition probability matrices C_i, Y_i in block forms as $C_i = (C_i^{(r,r')})_{r,r'=\overline{0,N}}, Y_i = (Y_i^{(r,r')})_{r,r'=\overline{0,N}}$, where the blocks $C_i^{(r,r')}$, $Y_i^{(r,r')}$ correspond to the transitions of the number of busy servers from r to r' at Station 2.

Denote by $\delta_{r,r'}(t)$ the probability that during the time interval of the length t the number of busy servers at Station 2 decreases from r to r' conditioned on the fact that none arrive from Station 1.

For use in the sequel, we register the following probabilistic interpretations of the matrices.

The (ν,ν')th entry of the matrix $P(j,t)$ gives the probability that j customers arrive in the BMAP during the interval $(0,t]$ and the state of the BMAP at epoch t is ν' given $\nu_0 = \nu$.

The (ν,ν')th entry of the matrix $\int_0^\infty \delta_{r,r'}(t) P(j,t)\, dB(t)$ gives the probability that during the service time of a customer at Station 1, exactly j customers arrive, the number of busy servers at Station 2 decreases from r to r' and the BMAP has changed from ν to ν'.

The (ν,ν')th entry of the matrix $\int_0^\infty \delta_{r,r'}(t) e^{D_0 t} D_k\, dt$ gives the probability that at an arbitrary time instant with the number of busy servers at Station 2 equal to r, and the BMAP in state ν, the first batch to arrive is of size k; soon after that instant, the number of busy servers at Station 2 equals r' and the BMAP is in state ν'.

$F_{r,r'}(t)$ is the distribution function of the time interval during which the number of busy servers at Station 2 decreases from r to r' conditioned on the fact that none arrive to this station. Then the (ν,ν')th entry of the matrix $\int_0^\infty P(j,t)dF_{r,r'}(t)$ defines the probability that exactly j customers arrive with the BMAP moving from ν to ν' and that the number of busy servers at Station 2 decreases from r to r' during that time interval.

From the above probabilistic interpretations, analyzing the one-step transitions of the chain ξ_n with a careful analysis of the service completion epochs, in which the customers may get lost due to lack of servers or wait (and thus block Station 1) until enough servers are available, we obtain the following expressions for the matrices $C_i^{(r,r')}, Y_i^{(r,r')},\ i \ge 0$:

$$\begin{aligned}
C_i^{(r,r')} &= \sum_{m=0}^{N-r} q_m \sum_{l=r'}^{r+m} \int_0^\infty \delta_{r+m,l}(t) e^{D_0 t} dt \sum_{k=1}^{i+1} D_k \int_0^\infty \delta_{l,r'}(t) P(i-k+1,t) dB(t) \\
&\quad + \gamma \sum_{m=N-r+1}^{N} q_m \sum_{l=r'}^{r} \int_0^\infty \delta_{r,l}(t) e^{D_0 t} dt \sum_{k=1}^{i+1} D_k \int_0^\infty \delta_{l,r'}(t) P(i-k+1,t) dB(t) \\
&\quad + (1-\gamma) \sum_{m=N-r+1}^{N} q_m \Bigg[\int_0^\infty e^{D_0 t} dF_{r,N-m}(t) \sum_{l=r'}^{N} \int_0^\infty \delta_{N,l}(t) e^{D_0 t} dt \\
&\qquad\qquad \times \sum_{k=1}^{i+1} D_k \int_0^\infty \delta_{l,r'}(t) P(i-k+1,t) dB(t) + \sum_{k=1}^{i+1} \int_0^\infty P(k,t) dF_{r,N-m}(t) \\
&\qquad\qquad \times \int_0^\infty \delta_{N,r'}(t) P(i-k+1,t) dB(t) \Bigg], \\
Y_i^{(r,r')} &= \sum_{m=0}^{N-r} q_m \int_0^\infty P(i,t) \delta_{r+m,r'}(t) dB(t) \\
&\quad + \sum_{m=N-r+1}^{N} q_m \Bigg[\gamma \int_0^\infty \delta_{r,r'}(t) P(i,t) dB(t) \\
&\qquad\qquad + (1-\gamma) \sum_{k=0}^{i} \int_0^\infty P(k,t) dF_{r,N-m}(t) \int_0^\infty \delta_{N,r'}(t) P(i-k,t) dB(t) \Bigg].
\end{aligned} \tag{3.3}$$

In order to arrive at equations (3.2) from (3.3), we use the matrix notations introduced above and the relation: $(\delta_{r,r'}(t))_{r,r'=\overline{0,N}} = e^{\Delta t}$, which follows from the fact that under the case when none arrive to Station 2, the process r_t governing the number of busy servers at this station is Markovian with generator Δ. □

It is easy to see that the Markov chain ξ_n belongs to the class of $M/G/1$ type Markov chains, see [16]. We can use this fact to derive the ergodicity condition and calculate the stationary distribution of the chain.

Let $C(z) = \sum_{i=0}^\infty C_i z^i, Y(z) = \sum_{i=0}^\infty Y_i z^i,\ |z| \le 1$, be the generating functions of the transition probability matrices C_i and $Y_i,\ i \ge 0$.

Corollary 3.2. *The matrix generating functions* $C(z)$,$\Upsilon(z)$ *can be written as*

$$C(z) = \frac{1}{z}\left[-\widehat{Q}(\Delta \oplus D_0)^{-1}\left(\widetilde{D}(z) - \widetilde{D}_0\right) + (1-\gamma)(F(z) - F_0)\widetilde{Q}_3\right]\Omega(z), \tag{3.4}$$

$$\Upsilon(z) = \left[\widetilde{Q}_1 + \gamma\widetilde{Q}_2 + (1-\gamma)F(z)\widetilde{Q}_3\right]\Omega(z), \tag{3.5}$$

where

$$\Omega(z) = \sum_{n=0}^{\infty}\Omega_n z^n = \int_0^{\infty} e^{\Delta t} \otimes e^{D(z)t} dB(t), \qquad F(z) = \sum_{n=0}^{\infty}F_n z^n = \int_0^{\infty} dF(t) \otimes e^{D(z)t}. \tag{3.6}$$

Theorem 3.3. *The necessary and sufficient condition for ergodicity of the Markov chain* ξ_n, $n \geq 1$, *is the fulfillment of the inequality*

$$\rho = \lambda\left[b_1 + (1-\gamma)\sum_{r=1}^{N}\vartheta_r \sum_{m=N-r+1}^{N} q_m \sum_{l=N-m+1}^{r} (l\mu)^{-1}\right] < 1. \tag{3.7}$$

Here $(\vartheta_1, \ldots, \vartheta_N)$ *is a part of the vector* $\boldsymbol{\vartheta} = (\vartheta_0, \ldots, \vartheta_N)$, *which is the unique solution to the system*

$$\boldsymbol{\vartheta} Q B^*(0) = \boldsymbol{\vartheta}, \qquad \boldsymbol{\vartheta}\mathbf{e} = 1, \tag{3.8}$$

where $B^*(s) = \int_0^{\infty} e^{-st}e^{\Delta t}dB(t)$.

Proof. It can be verified that the matrix $\Upsilon(1)$ is irreducible. Hence, from [16], the necessary and sufficient condition for ergodicity of the chain ξ_n is the fulfillment of the inequality

$$\mathbf{x}\Upsilon'(1)\mathbf{e} < 1, \tag{3.9}$$

where the vector $\mathbf{x}$ is the unique solution of the system

$$\mathbf{x}\Upsilon(1) = \mathbf{x}, \qquad \mathbf{x}\mathbf{e} = 1. \tag{3.10}$$

The theorem will be proven if we show that inequality (3.9) is equivalent to inequality (3.7).

Let the vector $\mathbf{x}$ be of the form

$$\mathbf{x} = \boldsymbol{\vartheta} \otimes \boldsymbol{\theta}. \tag{3.11}$$

By the direct substitution into the system (3.10), where $\Upsilon(1)$ is calculated using (3.5), we verify that such a vector provides the unique solution of this system. Differentiating (3.5) at

Table 1: The value of the system load ρ for different value of the mean service time and service time variation.

	$c_{\text{var}} = 0$	$c_{\text{var}} = 1$	$c_{\text{var}} = 5$	$c_{\text{var}} = 9.95$
$b_1 = 0.1$	0.42754	0.43016	0.45557	0.47940
$b_1 = 0.2$	0.47460	0.48228	0.53677	0.57784
$b_1 = 0.3$	0.53173	0.54515	0.62243	0.67725
$b_1 = 0.4$	0.59698	0.61590	0.71010	0.77695
$b_1 = 0.5$	0.66879	0.69247	0.79902	0.87678
$b_1 = 0.6$	0.74579	0.77339	0.88883	0.97667
$b_1 = 0.7$	0.82687	0.85763	0.97937	**1.07663**
$b_1 = 0.8$	0.91116	0.94440	**1.07051**	**1.17655**

the point $z = 1$ and substituting the resulting expression for $Y'(1)$ and the vector $\mathbf{x}$ of form (3.11) into the inequality (3.9), we get

$$\rho = \lambda \left[b_1 + (1-\gamma)\boldsymbol{\vartheta} \int_0^\infty t dF(t) Q_3 \mathbf{e} \right] < 1. \tag{3.12}$$

The stated expression in (3.7) follows from (3.12) and the expression for $\int_0^\infty t dF(t)$ given in (C.3)-(C.4) (see Appendix C). □

Remark 3.4. The inequality (3.7) is intuitively clear on noting that the vector $\boldsymbol{\vartheta}$ gives the stationary distribution of the number of busy servers at Station 2 at the service completion epochs at Station 1 given the latter station works non-stop. Then $(1-\gamma)\sum_{r=1}^{N}\vartheta_r \sum_{m=N-r+1}^{N} q_m \sum_{l=N-m+1}^{r}(l\mu)^{-1}$ defines the average blocking time of Station 1 under overload condition and ρ is the system load.

Remark 3.5. In a majority of queueing systems, the system load depends only on the first moment of the service time distribution. In the model under study, the value of ρ depends not only on the first moment b_1 of the service time distribution at Station 1, but also on the shape of this distribution. In particular, ρ depends on variance of the service time. This fact is illustrated in Table 1 in Section 5.

In what follows, we assume that the inequality (3.7) holds true.

Denote the stationary state probabilities of the Markov chain $\xi_n = \{i_n, r_n, \nu_n\}$ by $\pi(i,r,\nu)$, $i \geq 0$, $r = \overline{0,N}$, $\nu = \overline{0,W}$. Introduce the notation for the row vectors of these probabilities

$$\boldsymbol{\pi}(i,r) = (\pi(i,r,0), \pi(i,r,1), \ldots, \pi(i,r,W)), \qquad \boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1), \ldots, \boldsymbol{\pi}(i,N)), \qquad i \geq 0. \tag{3.13}$$

Let also $\boldsymbol{\Pi}(z) = \sum_{i=0}^{\infty} \boldsymbol{\pi}_i z^i$, $|z| \leq 1$, be the vector generating function of vectors $\boldsymbol{\pi}_i$, $i \geq 0$. To compute these vectors as well as the vectors $\boldsymbol{\Pi}(1)$ and $\boldsymbol{\Pi}'(1)$, known algorithms, see, for example, [16], can be applied.

Once the stationary distribution has been computed, we can calculate some key performance measures of the system as follows.

(i) The mean number of customers at Station 1 at the service completion epochs $L = \mathbf{\Pi}'(1)\mathbf{e}$.

(ii) The vector of the stationary distribution of the number of busy servers at Station 2 at the service completion epoch at Station 1: $\mathbf{r} = \mathbf{\Pi}(1)(I_{N+1} \otimes \mathbf{e}_{\overline{W}})$.

(iii) The mean number of busy servers at Station 2 at the service completion epoch at Station 1

$$N_{\text{busy}} = \mathbf{r}\ \text{diag}\left\{ r,\ r = \overline{0, N} \right\}\mathbf{e}. \tag{3.14}$$

(iv) The probability that an arbitrary customer leaves the system or causes the blocking of the server at Station 1

$$P_{\text{loss}} = \gamma \mathbf{\Pi}(1)\widetilde{Q}_2\mathbf{e}, \qquad P_{\text{block}} = (1-\gamma)\mathbf{\Pi}(1)\widetilde{Q}_2\mathbf{e}. \tag{3.15}$$

(v) The probability that the server of Station 1 is idle at an arbitrary time $p_{\text{idle}} = \tau^{-1}\boldsymbol{\pi}_0\widehat{Q}(-\widetilde{D}_0)^{-1}\mathbf{e}$, where τ is the mean interdeparture time at Station 1,

$$\tau = b_1 + \boldsymbol{\pi}_0\widehat{Q}\left(-\widetilde{D}_0\right)^{-1}\mathbf{e} - (1-\gamma)\mathbf{\Pi}(1)(I_{N+1} \otimes \mathbf{e})F^{(1)}Q_3\mathbf{e}, \tag{3.16}$$

the matrix $F^{(1)}$ is defined by formula (C.3) below.

(vi) The probability that the server of Station 1 processes a customer at an arbitrary time $p_{\text{serve}} = \tau^{-1}b_1$.

(vii) The probability that the server of Station 1 is blocked at an arbitrary time $p_{\text{block}} = 1 - p_{\text{idle}} - p_{\text{serve}}$.

4. Stationary Distribution of the Sojourn Time

4.1. The Virtual Sojourn Time

The virtual sojourn time in the system consists of the virtual sojourn time at Station 1 and the sojourn time at Station 2. We assume that customers are served according to FIFO (first-in-first-out) discipline.

For use in the sequel, we define the *generalized service time* of an arbitrary customer as the service time of this customer by the first server and the possible blocking time of the server by the previous customer.

4.1.1. The Virtual Sojourn Time at Station 1

The virtual sojourn time at Station 1 consists of (i) the *residual time* from an arbitrary time instant (associated with virtual customer arrival) to the next service completion epoch at Station 1 (ii) the *generalized service times* of customers staying in the queue at an arbitrary time, and (iii) the *generalized service time* of the virtual customer.

First, we study the *residual time*. To this end, we consider the process $\chi_t = \{i_t,\ m_t,\ r_t,\ \nu_t,\ \tilde{v}_t\},\ t \geq 0$, whose components are defined as follows: i_t is the number of customers in Station 1 (including the blocked customer, if any), m_t takes values 0, 1, 2, respectively, based on the server at Station 1 is idle, busy, or blocked at time t, ν_t is the state of the BMAP, r_t is the number of busy servers at Station 2 just before the service completion epoch following the time t, $\tilde{v}_t$ is the residual time from t to that service completion epoch.

Using the definition of semiregenerative processes given in [17], it can be verified that the process χ_t is a semi-regenerative one with the embedded Markov renewal process $\{\xi_n, t_n\}$, $n \geq 1$. Let

$$\begin{aligned}
&\widetilde{V}(i,m,r,\nu,x)\\
&\quad = \lim_{t\to\infty} P\{i_t = i, m_t = m, r_t = r, \nu_t = \nu, \tilde{v}_t < x\}, \quad i \geq 0,\ m = \overline{0,2},\ r = \overline{0,N},\ \nu = \overline{0,W},\ x \geq 0,
\end{aligned} \tag{4.1}$$

be the stationary distribution of the process $\chi_t,\ t \geq 0$.

From [17], the limits in (4.1) exist if the process $\{\xi_n, t_n\},\ n \geq 1$, is irreducible aperiodic recurrent and the value τ of the mean inter-departure time at Station 1 (given by (3.16)) is finite. All these conditions hold if inequality (3.7) is satisfied.

Let $\widetilde{\mathbf{V}}(i,m,x)$ be the row vector of the steady state probabilities $\widetilde{V}(i,m,r,\nu,x)$ arranged according to the lexicographic order of the components (r,ν), and let $\tilde{\mathbf{v}}(i,m,s)$ be the corresponding vector of the Laplace-Stieltjes transforms, that is, let $\tilde{\mathbf{v}}(i,m,s) = \int_0^\infty e^{-sx} d\widetilde{\mathbf{V}}(i,m,x),\ i \geq 0,\ m = \overline{0,2}$.

Lemma 4.1. *The vector Laplace-Stieltjes transforms $\tilde{\mathbf{v}}(i,m,s)$ are calculated by*

$$\tilde{\mathbf{v}}(0,1,s) = \mathbf{0},\ \tilde{\mathbf{v}}(i,0,s) = \mathbf{0}, \quad i > 0, \quad \tilde{\mathbf{v}}(0,2,s) = \mathbf{0}, \tag{4.2}$$

$$\tilde{\mathbf{v}}(0,0,s) = -\tau^{-1}\boldsymbol{\pi}_0\widehat{Q}(\Delta \oplus D_0)^{-1}\left[B^*(s) \otimes I_{\overline{W}}\right], \tag{4.3}$$

$$\begin{aligned}
\tilde{\mathbf{v}}(i,1,s) = \tau^{-1}\Bigg\{ &\boldsymbol{\pi}_0 \sum_{k=1}^{i}\left[-\widehat{Q}(\Delta \oplus D_0)^{-1}\widetilde{D}_k + (1-\gamma)F_k\widetilde{Q}_3\right]\\
&\times \int_0^\infty \left(e^{\Delta u} \otimes I_{\overline{W}}\right) \int_0^u I_{N+1} \otimes P(i-k,y)e^{-s(u-y)}dy\, dB(u)\\
&+ \sum_{j=1}^{i}\boldsymbol{\pi}_j\Bigg[\left(\widetilde{Q}_1 + \gamma\widetilde{Q}_2\right)\int_0^\infty \left(e^{\Delta u} \otimes I_{\overline{W}}\right)\int_0^u I_{N+1} \otimes P(i-j,y)e^{-s(u-y)}dy\, dB(u)\\
&+(1-\gamma)\sum_{k=0}^{i-j}F_k\widetilde{Q}_3\int_0^\infty e^{\Delta u} \otimes P(i-k-j,y)e^{-s(u-y)}dy\, dB(u)\Bigg]\Bigg\},
\end{aligned} \tag{4.4}$$

$$\tilde{\mathbf{v}}(i,2,s) = \tau^{-1}(1-\gamma)\sum_{j=0}^{i-1}\boldsymbol{\pi}_j\left(\int_0^\infty dF(u)\otimes I_{\overline{W}}\right)\tilde{Q}_3 \\ \times\int_0^u e^{-s(u-y)}\left[I_{N+1}\otimes P(i-j-1,y)\right]dy\left[B^*(s)\otimes I_{\overline{W}}\right], \quad i>0. \tag{4.5}$$

Proof. Let $\kappa_j^{(i,m,x,t)}(r,\nu;\ r',\nu')$ denote the conditional probability that, given time 0 is an instant of the service completion at Station 1 and the embedded Markov chain ξ_n is in the state (j,r,ν) at that time, the next service completion epoch at Station 1 occurs later than t, the discrete components of the process χ_t take values (i,m,r',ν') at time t and the continuous-time component $\tilde{v}_t < x$.

Let us arrange the probabilities $\kappa_j^{(i,m,x,t)}(r,\nu;r',\nu')$, for fixed values $i,\ j,\ m,$ according to the lexicographic order of the states $(r,\nu;r',\nu')$ and form the square matrices

$$\tilde{K}_j(i,m,x,t) = \left(\kappa_j^{(i,m,x,t)}(r,\nu;r',\nu')\right)_{\nu,\nu'=\overline{0,W};r,r'=\overline{0,N}}. \tag{4.6}$$

Then, using the ergodic theorem for semi-regenerative processes, (see [17, Theorem 6.12]), the probability vectors $\tilde{\mathbf{V}}(i,m,x)$ can be related to the stationary distribution $\boldsymbol{\pi}_j$, $j\ge 0$, of the embedded Markov chain $\xi_n, n\ge 1$, by

$$\tilde{\mathbf{V}}(i,m,x) = \tau^{-1}\sum_{j=0}^{\infty}\boldsymbol{\pi}_j\int_0^\infty \tilde{K}_j(i,m,x,t)dt, \quad i\ge 0,\ m=\overline{0,2}. \tag{4.7}$$

The corresponding vector Laplace-Stieltjes transforms $\tilde{\mathbf{v}}(i,m,s)$ are defined by

$$\tilde{\mathbf{v}}(i,m,s) = \tau^{-1}\sum_{j=0}^{\infty}\boldsymbol{\pi}_j\int_0^\infty \tilde{K}_j^*(i,m,s,t)dt, \quad i\ge 0,\ m=\overline{0,2}, \tag{4.8}$$

where $\tilde{K}_j^*(i,m,s,t) = \int_0^\infty e^{-sx}d\tilde{K}_j(i,m,x,t)$.

From (4.8), formulas (4.2) follow immediately when we note that $\tilde{K}_j^*(i,m,s,t)=0$ for the range of arguments $\{j\ge 0, i=0, m=1,2\}$ and $\{j\ge 0, i>0, m=0\}$.

Let $m=1$. The lengthy but straightforward expressions for the matrices $\tilde{K}_j^*(i,1,s,t)$, $i>0$, are presented in Appendix A. Substituting these expressions into (4.8) and after routine algebraic manipulations including rearranging the order of integration, we get formula (4.4) for the vectors $\tilde{\mathbf{v}}(i,1,s)$, $i>0$. Similar calculations yield (4.3) and (4.5). □

Further, we study the *generalized service time* distribution at Station 1.

Let $\widehat{B}(x)$ be the matrix distribution function of *generalized service time*. More specifically, let $\widehat{B}(x) = (\widehat{B}(x)_{r,r'})_{r,r'=\overline{0,N}}$, where $\widehat{B}(x)_{r,r'} = P\{t_{n+1} - t_n < x, r_{n+1} = r' \mid r_n = r, i_n \neq 0\}$. Denote $\mathcal{B}(s) = \int_0^\infty e^{-st} d\widehat{B}(t)$.

Lemma 4.2. *The matrix Laplace-Stieltjes transform of the generalized service time distribution at Station 1 is calculated as*

$$\mathcal{B}(s) = \left[Q_1 + \gamma Q_2 + (1-\gamma) F^*(s) Q_3\right] B^*(s), \tag{4.9}$$

where $F^*(s) = \int_0^\infty e^{-st} dF(t)$.

Proof. To prove, we need to analyze the structure of the generalized service time. The generalized service time of a tagged customer is just the service time of the customer at Station 1 if the previous customer did not block the server of this station. In this case, the matrix Laplace-Stieltjes transform of the generalized service time distribution is calculated by $(Q_1 + \gamma Q_2) B^*(s)$. However, when blocking occurs, the generalized service time consists of the time during which the server is blocked by the previous customer and the service time of the tagged customer. The corresponding Laplace-Stieltjes transform is defined by $(1-\gamma) \int_0^\infty e^{-st} dF(t) Q_3 B^*(s)$. The stated result (4.9) follows immediately. □

Now we are ready to derive the equation for the vector Laplace-Stieltjes transform $\mathbf{v}_1(s)$ of the distribution of the virtual sojourn time at Station 1. Let $v_1(r, \nu, x)$ be the probability that, at an arbitrary epoch, the BMAP is in state ν, the virtual sojourn time at Station 1 is less than x, and the number of busy servers at Station 2 just before the end of the virtual sojourn time is r. Then $\mathbf{v}_1(s)$ is defined as a vector of Laplace-Stieltjes transforms $v_1(r, \nu, s) = \int_0^\infty e^{-sx} dv_1(r, \nu, x)$ written in lexicographic order.

Theorem 4.3. *The vector Laplace-Stieltjes transform* $\mathbf{v}_1(s)$ *satisfies the equation*

$$\mathbf{v}_1(s) A(s) = \boldsymbol{\pi}_0 \Phi(s), \tag{4.10}$$

where

$$A(s) = sI + \sum_{r=0}^{\infty} \mathcal{B}^r(s) \otimes D_r, \qquad \Phi(s) = \tau^{-1} \widehat{Q} (\Delta \oplus D_0)^{-1} \left(\Delta \otimes I_{\overline{W}} - sI\right) \left[B^*(s) \otimes I_{\overline{W}}\right]. \tag{4.11}$$

Proof. As mentioned above, the virtual sojourn time at Station 1 consists of the residual time from an arbitrary time t to the next service completion epoch, the generalized service times of customers that await for a service at time t, and the generalized service time of the virtual customer.

Taking into account the structure of the virtual sojourn time and using the law of total probability, we express the vector Laplace-Stieltjes transform $\mathbf{v}_1(s)$ as follows:

$$\mathbf{v}_1(s) = \tilde{\mathbf{v}}(0,0,s) + \sum_{i=1}^{\infty} \tilde{\mathbf{v}}(i,1,s) \left[\mathcal{B}^i(s) \otimes I_{\overline{W}}\right] + \sum_{i=1}^{\infty} \tilde{\mathbf{v}}(i,2,s) \left[\mathcal{B}^{i-1}(s) \otimes I_{\overline{W}}\right]. \tag{4.12}$$

Further, we multiply (4.12) by the matrix $sI+\sum_{r=0}^{\infty}\mathcal{B}^r(s)\otimes D_r$, and, after some laborious calculations, get

$$\begin{aligned}
&\mathbf{v}_1(s)\left(sI+\sum_{r=0}^{\infty}\mathcal{B}^r(s)\otimes D_r\right)\\
&\quad=\tau^{-1}\left\{\boldsymbol{\pi}_0\sum_{i=0}^{\infty}\left[C_i+\sum_{j=1}^{i+1}\boldsymbol{\pi}_jY_{i-j+1}\right]\left[\mathcal{B}^{i+1}(s)\otimes I_{\overline{W}}\right]\right.\\
&\qquad\left.+\boldsymbol{\pi}_0\widehat{Q}\left[-(\Delta\oplus D_0)^{-1}\left(sI+\widetilde{D}_0\right)+I\right]\left[B^*(s)\otimes I_{\overline{W}}\right]-\sum_{j=0}^{\infty}\boldsymbol{\pi}_j\left[\mathcal{B}^{j+1}(s)\otimes I_{\overline{W}}\right]\right\}.
\end{aligned}\tag{4.13}$$

Multiplying the balance equations for stationary probability vectors $\boldsymbol{\pi}_i$ of the form

$$\boldsymbol{\pi}_i=\boldsymbol{\pi}_0C_i+\sum_{l=1}^{i+1}\boldsymbol{\pi}_lY_{i-l+1},\quad i\geq 0,\tag{4.14}$$

by $\mathcal{B}^{i+1}(s)\otimes I_{\overline{W}}$ and summing over i we obtain

$$\sum_{i=0}^{\infty}\boldsymbol{\pi}_i\left[\mathcal{B}^{i+1}(s)\otimes I_{\overline{W}}\right]=\boldsymbol{\pi}_0\sum_{i=0}^{\infty}C_i\left[\mathcal{B}^{i+1}(s)\otimes I_{\overline{W}}\right]+\sum_{i=0}^{\infty}\sum_{j=1}^{i+1}\boldsymbol{\pi}_jY_{i-j+1}\left[\mathcal{B}^{i+1}(s)\otimes I_{\overline{W}}\right].\tag{4.15}$$

Using (4.15) to simplify equation (4.13), we obtain (4.10). □

4.1.2. The Sojourn Time at Station 2

Let $\mathbf{v}_2(s)$ be the column vector of the Laplace-Stieltjes transforms of the conditional sojourn time distributions at Station 2. The rth entry of this vector is the Laplace-Stieltjes transform of the sojourn time distribution of a customer at Station 2 given that the number of busy servers is equal to r just before the end of the sojourn time of this customer at Station 1.

Lemma 4.4. *The vector Laplace-Stieltjes transform of the sojourn time distribution at Station 2 is given by*

$$\mathbf{v}_2(s)=\left[Q_4(F^*(s)+I)\widehat{I}+\gamma Q_2+(1-\gamma)F^*(s)\operatorname{diag}\{f_{r,0}(s),r=N,N-1,\ldots,0\}Q_3\right]\mathbf{e}.\tag{4.16}$$

Proof. The sojourn time of a customer who requires m servers at Station 2 consists of:

(i) the service time of the customer when at least m servers are available at the time of the request;

(ii) zero time, if the required number of servers is not available and the customer leaves the system;

(iii) the blocking time and the service time of a customer when the required number of servers is not available and the customer awaits the release of sufficiently many servers.

Note that we assume that the service of type m is performed by m servers independently of each other and finishes when all m servers complete the service. The distribution of this service time is defined by the Laplace-Stieltjes transform $f_{m,0}(s)$, $m = \overline{1,N}$.

Taking into account this fact together with (i)–(iii) and using the matrix notation, we obtain expression (4.16) for the vector $\mathbf{v}_2(s)$. In the expression, the first summand corresponds to the case (i), the second and the third summands give the Laplace-Stieltjes transform under study in the cases (ii) and (iii), respectively. □

4.1.3. The Virtual Sojourn Time in the System

Theorem 4.5. *The Laplace-Stieltjes transform of the virtual sojourn time distribution in the system is given by*

$$v(s) = \mathbf{v}_1(s)\left(I_{N+1} \otimes \mathbf{e}_{\overline{W}}\right)\mathbf{v}_2(s), \tag{4.17}$$

where the vectors $\mathbf{v}_1(s)$ *and* $\mathbf{v}_2(s)$*, respectively, are as given in* (4.10) *and* (4.16)*.*

Proof. Formula (4.17) readily follows from the structure of the virtual sojourn time in the system which consists of the virtual sojourn time at Station 1 and the sojourn time at Station 2. □

4.2. The Actual Sojourn Time

Let $v_1^{(a)}(s)$ and $v^{(a)}(s)$ be the Laplace-Stieltjes transforms of the distribution of the actual sojourn time at Station 1 and in the whole system.

Theorem 4.6. *The Laplace-Stieltjes transform of the actual sojourn time distribution at Station 1 is calculated as follows:*

$$v_1^{(a)}(s) = \lambda^{-1}\mathbf{v}_1(s)\sum_{k=0}^{\infty}\left[\mathcal{B}^k(s)(\mathcal{B}(s) - I)^{-1} \otimes D_k\right]\mathbf{e}. \tag{4.18}$$

Proof. The actual sojourn time at Station 1 of an arbitrary-tagged customer, who arrived in a group of size k and placed at the jth position within the group, consists of (a) the actual sojourn time at Station 1 of the first customer in the group, which coincides with the virtual sojourn time at Station 1; (b) the generalized service times at Station 1 of the $j-2$ customers of the group who arrived with the tagged customer; (c) the generalized service time of the tagged customer at Station 1.

The vector Laplace-Stieltjes transform of the sojourn time distribution at Station 1 of the first customer of the k-size group that contains the tagged customer is evidently given by the vector $\mathbf{v}_1(s)(I_{N+1} \otimes kD_k\mathbf{e}/\lambda)$.

Assuming that an arbitrary customer arriving in a group of size k is placed on the jth position with probability $1/k$ and using the law of total probability, we immediately obtain the following expression:

$$v_1^{(a)}(s) = \sum_{k=1}^{\infty} \mathbf{v}_1(s)\left(I_{N+1} \otimes \frac{kD_k\mathbf{e}}{\lambda}\right)\sum_{j=1}^{k}\frac{1}{k}\mathcal{B}^{j-1}(s)\mathbf{e}. \tag{4.19}$$

After some algebraic manipulations (4.19) is reduced to (4.18). □

Corollary 4.7. *The Laplace-Stieltjes transform of the actual sojourn time distribution in the whole system is calculated as follows:*

$$v^{(a)}(s) = \lambda^{-1}\mathbf{v}_1(s)\sum_{k=0}^{\infty}\left[\mathcal{B}^k(s)(\mathcal{B}(s) - I)^{-1} \otimes D_k\right]\left(I_{N+1} \otimes \mathbf{e}_{\overline{W}}\right)\mathbf{v}_2(s). \tag{4.20}$$

4.3. Moments of the Sojourn Time Distribution

The formulas for the moments of the virtual sojourn time distribution can be obtained by differentiating the expression in (4.10) at the point $s = 0$. This requires the calculation of the derivatives of $\mathbf{v}_1(s)$ at the point $s = 0$. However, the matrix $A(s)$ in (4.10) is singular at the point $s = 0$ and calculation of the derivatives at this point is the nontrivial task. The results given below allow one to develop a procedure for calculating the required derivatives.

We will use notation $\mathbf{v}_1^{(m)}(s)$ for the mth derivative of the vector $\mathbf{v}_1(s)$, $m \geq 1$, and set $\mathbf{v}_1^{(0)}(s) = \mathbf{v}_1(s)$. Similar notations will be used for other functions of s.

Theorem 4.8. *Let* $\int_0^\infty t^m dB(t) < \infty$, $m = \overline{1, M+1}$, *where M is an arbitrary positive integer. Then the vectors* $\mathbf{v}_1^{(m)}(0)$, $m = \overline{1, M}$, *are computed recursively by*

$$\begin{aligned}\mathbf{v}_1^{(m)}(0) &= \left[\left(\boldsymbol{\pi}_0\Phi^{(m)}(0) - \sum_{l=0}^{m-1}\binom{m}{l}\mathbf{v}_1^{(l)}(0)A^{(m-l)}(0)\right)\tilde{I}\right. \\ &\quad \left. + \frac{1}{m+1}\left(\boldsymbol{\pi}_0\Phi^{(m+1)}(0) - \sum_{l=0}^{m-1}\binom{m+1}{l}\mathbf{v}_1^{(l)}(0)A^{(m+1-l)}(0)\right)\mathbf{e}\hat{\mathbf{e}}\right]\tilde{A}^{-1},\end{aligned} \tag{4.21}$$

with the initial condition

$$\mathbf{v}_1^{(0)}(0) = \mathbf{v}_1(0) = \left[\tau^{-1}\boldsymbol{\pi}_0\hat{Q}(\Delta \oplus D_0)^{-1}\left(\Delta B^*(0) \otimes I_{\overline{W}}\right)\tilde{I} + p_{\text{idle}}\,\hat{\mathbf{e}}\right]\tilde{A}^{-1}, \tag{4.22}$$

where $\tilde{A} = A(0)\tilde{I} + A'(0)\mathbf{e}\hat{\mathbf{e}}$.

Proof of the theorem is presented in Appendix B. In what follows we assume that $\int_0^\infty t^k dB(t) < \infty$, $k = \overline{1,2}$.

Corollary 4.9. *The mean virtual sojourn time at Station 1 is given by*

$$\begin{aligned}\overline{v}_1 = \Big\{ &\Big[\tau^{-1}\boldsymbol{\pi}_0\widehat{Q}(\Delta \oplus D_0)^{-1}\Big(\Big(B^*(0) - \Delta B^{*'}(0)\Big)\otimes I_{\overline{W}}\Big) + \mathbf{v}_1(0)A'(0)\Big]\widetilde{I} \\ &+\Big[p_{idle}b_1 + \frac{1}{2}\mathbf{v}_1(0)A''(0)\mathbf{e}\Big]\widehat{\mathbf{e}}\Big\}\widetilde{A}^{-1}\mathbf{e},\end{aligned} \tag{4.23}$$

where the vector $\mathbf{v}_1(0)$ *is defined by formula* (4.22).

Proof follows from the relation $\overline{v}_1 = -\mathbf{v}_1'(0)\mathbf{e}$ and formula (4.21).

Theorem 4.10. *The mean virtual sojourn time in the system is given by*

$$\overline{v} = \overline{v}_1 + \mathbf{v}_1(0)\left(I_{N+1} \otimes \mathbf{e}_{\overline{W}}\right)\overline{\mathbf{v}}_2, \tag{4.24}$$

where $\mathbf{v}_1(0)$ *and* $\overline{v}_1$ *are given in* (4.22) *and* (4.23), *respectively, and* $\overline{\mathbf{v}}_2$ *is a vector of conditional means of the sojourn time at Station 2,*

$$\overline{\mathbf{v}}_2 = -\left\{Q_4F^{(1)}\widehat{\mathbf{e}}^T + (1-\gamma)\left[F^{(1)} - E\,\mathrm{diag}\left\{\sum_{l=1}^{r}(l\mu)^{-1},\ r = N, N-1, \dots, 0\right\}\right]Q_3\mathbf{e}\right\}, \tag{4.25}$$

where $F^{(1)}$ *is given by formula* (C.3) *below.*

Proof. To calculate the value $\overline{v}$ we differentiate the expression in (4.17). Setting $s = 0$ and replacing the sign, we have that

$$\overline{v} = -\mathbf{v}_1'(0)\left(I_{N+1} \otimes \mathbf{e}_{\overline{W}}\right)\mathbf{v}_2(0) - \mathbf{v}_1(0)\left(I_{N+1} \otimes \mathbf{e}_{\overline{W}}\right)\mathbf{v}_2'(0). \tag{4.26}$$

Putting $s = 0$ in (4.16) we get $\mathbf{v}_2(0) = [Q_4(E + I)\widehat{I} + \gamma Q_2 + (1-\gamma)EQ_3]\mathbf{e} = \mathbf{e}$. This implies that the first term in the right-hand side of (4.26) is equal to $-\mathbf{v}_1'(0)\mathbf{e} = \overline{v}_1$. Using the relation $\overline{\mathbf{v}}_2 = -\mathbf{v}_2'(0)$ and differentiating (4.16) at the point $s = 0$, we readily verify that $\overline{\mathbf{v}}_2$ has form (4.25). This completes the proof. □

Theorem 4.11. *The mean actual sojourn time at Station 1 is given by*

$$\overline{v}_1^{(a)} = -\lambda^{-1}\left\{\mathbf{v}_1'(0)\left(\mathbf{e} \otimes \sum_{k=1}^{\infty}kD_k\mathbf{e}\right) + \mathbf{v}_1(0)\sum_{k=1}^{\infty}(I_{N+1} \otimes D_k\mathbf{e})\left[\sum_{n=1}^{k-1}\sum_{l=0}^{n-1}\mathcal{B}^l(0)\mathcal{B}'(0)\mathbf{e}\right]\right\}. \tag{4.27}$$

Proof of the theorem follows from the relation $\overline{v}_1^{(a)} = -dv_1^{(a)}(s)/ds|_{s=0}$ and formula (4.19).

Corollary 4.12. *The mean actual sojourn time in the system is given by*

$$\overline{v}_a = \overline{v}_1^{(a)} + \lambda^{-1}\mathbf{v}_1(0)\sum_{k=1}^{\infty}(I_{N+1} \otimes D_k\mathbf{e})\sum_{n=0}^{k-1}\mathcal{B}^n(0)\overline{\mathbf{v}}_2. \tag{4.28}$$

5. Numerical Examples

In this section, we demonstrate feasibility of the algorithms developed here and show numerically some interesting features of the system under consideration.

Experiment 1. In this experiment, we investigate the impact of coefficient of variation in the service process at Station 1 on the main performance measures of the system.

To this end, we consider four service processes with the same mean service time $b_1 = 0.1$, but different values for the coefficient of variation, c_{var}. The first process is coded as D and corresponds to the deterministic service time distribution. The second process is coded as M and corresponds to the exponential service time. The third and the fourth service processes are coded as $HM_2^{(1)}$, $HM_2^{(2)}$ and correspond to hyperexponential service time distributions of order 2. These distributions are defined by the mixing probabilities $(0.05,\ 0.95)$ and the intensities $0.62025,\ 48.9998$ in the case of $HM_2^{(1)}$ and $(0.98,\ 0.02)$ and the intensities 10000, 0.2 in the case of $HM_2^{(2)}$. The coefficients of variation of processes D, M, $HM_2^{(1)}$, and $HM_2^{(2)}$ are, respectively, equal to $0, 1, 5, 9.95$.

The input process is defined by the matrices

$$D_0 = \begin{pmatrix} -15.7327 & 0.6062 & 0.5924 \\ 0.5178 & -2.2897 & 0.4679 \\ 0.5971 & 0.5653 & -1.9597 \end{pmatrix}, \qquad D = \begin{pmatrix} 14.1502 & 0.3021 & 0.0818 \\ 0.1071 & 1.032 & 0.1646 \\ 0.0858 & 0.1979 & 0.5136 \end{pmatrix}. \tag{5.1}$$

The matrices $D_k,\ k = \overline{1,5}$, are calculated as follows: $D_k = Dh^{k-1}(1-h)/(1-h^5)$, where $h = 0.8$. Then we normalize the matrices $D_k, k = \overline{0,5}$, so as to get the arrival rate $\lambda = 1$. This BMAP has coefficient of correlation $c_{\text{cor}} = 0.2$. The other parameters of the system are as follows: $N = 5,\ \mu = 0.8,\ \gamma = 0.5,\ q_0 = 0.1,\ q_1 = q_2 = 0.3,\ q_3 = q_4 = q_5 = 0.1$.

We vary the mean service time b_1 for all considered service processes in the interval $[0.1, 0.95]$ by scaling appropriately. The coefficients of variation do not change under such scaling. Note that, as it was mentioned in Remark 3.5 above, the system load ρ depends not only on the mean service time, but also on the variance of the service time. In Table 1, the value of ρ is given as function of b_1 and c_{var}. The values of ρ that exceed 1 are printed in bold face. The tandem queueing system is not stable for these values.

Figures 1 and 2 show the dependence of the main performance measures of the system on the value of the mean service time b_1 for service processes with different service time distribution. From these figures, it is very clear that the key performance measures of the system are very sensitive with respect to the service time variance. We also ran other examples, besides the one presented here, involving Erlangian and uniform distributions for the service time distribution. Since all these distributions have coefficient of variation in the range $(0, 1)$, the corresponding curves, as expected, were located between the two lower curves in Figures 1 and 2.

Experiment 2. Here we solve numerically the following optimization problem. Find an optimal choice for the number N of servers at Station 2 that will minimize the expected total

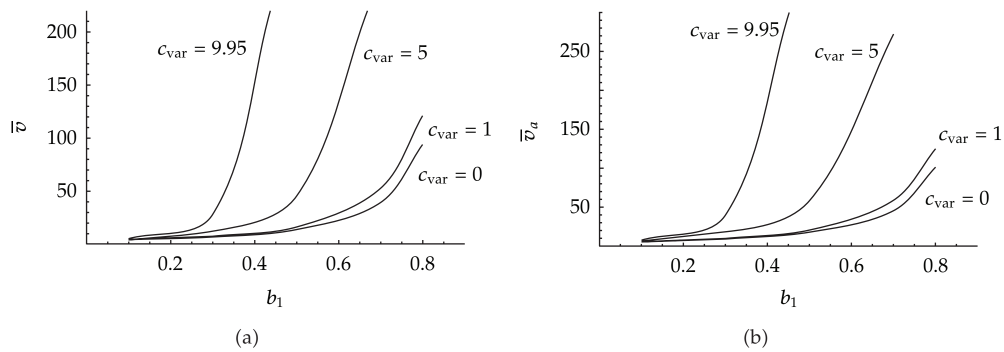

Figure 1: The mean virtual and actual sojourn time as functions of the mean service time for different service time distributions.

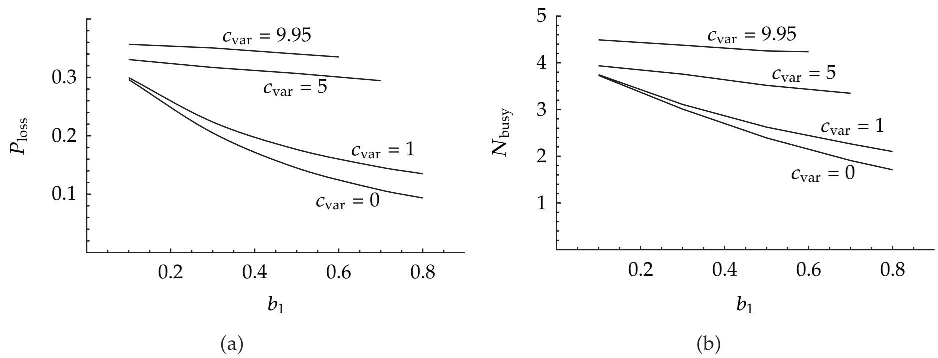

Figure 2: The loss probability and the mean number of busy servers at Station 2 as functions of the mean service time for different service time distributions.

cost per unit of time:

$$J = J(N) = aN + c_1 \lambda P_{\text{loss}} + c_2 \overline{v}^{(a)}, \tag{5.2}$$

where a is the cost of utilization per unit time of a server at Station 2 (maintenance cost), c_1 is the cost of a customer leaving the system after a service at Station 1 due to lack of required servers, and c_2 is the cost per unit of time of holding (the sojourn time) an arbitrary customer in the system (holding cost).

Using the MAP characterized by the matrices

$$D_0 = \begin{pmatrix} -6.74538 & 5.45412 \times 10^{-6} \\ 5.45412 \times 10^{-6} & -0.219455 \end{pmatrix}, \qquad D = \begin{pmatrix} 6.700685 & 0.044695 \\ 0.122427 & 0.097023 \end{pmatrix}, \tag{5.3}$$

we construct the BMAP with the matrices D_k, $k = \overline{0,5}$, similar to Experiment 1 and normalize so as to have $\lambda = 3$. The BMAP has the coefficient of correlation $c_{\text{cor}} = 0.2$ and the coefficient of variation $c_{\text{var}} = 3.5$.

Table 2: The value of the objective function for different number of servers and different service rate at Station 2.

	$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 4$	$\mu = 5$
$N = 1$	∞	457.0346	74.9351	51.0311	42.5146
$N = 2$	353.7378	48.6349	34.1130	27.0826	22.9120
$N = 3$	71.0480	34.6565	25.8441	**22.2656**	**20.6657**
$N = 4$	51.7822	**30.4068**	**25.8202**	24.6008	24.2393
$N = 5$	45.2556	31.2080	29.3925	29.1141	29.0630
$N = 6$	**43.0750**	34.6437	34.0980	34.0535	34.0486
$N = 7$	43.6867	39.0715	39.0533	39.0504	39.0476
$N = 8$	47.2564	44.0527	44.0514	44.0495	44.0474
$N = 9$	51.9854	49.0521	49.0511	49.0491	49.0473
$N = 10$	57.0056	55.0458	54.6001	54.0489	54.0472

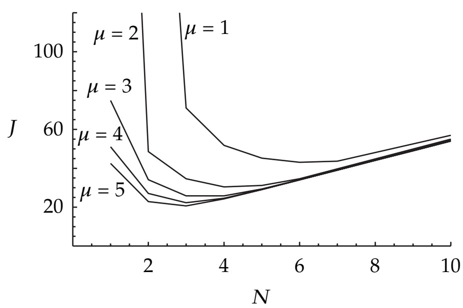

Figure 3: The objective function as a function of the number of servers in Station 2 for different service rates.

The service time distribution at Station 1 is assumed to be Erlang of order 3 with parameter 20. The probability γ is taken to be 0.5. The components of the vector $\mathbf{q}$ are: $q_0 = 0.1$, $q_1 = 0.9$, $q_m = 0$, $m = \overline{2, N}$. The various costs are taken as follows: $a = 5$, $c_1 = 50$, $c_2 = 3$.

The objective function, J, as a function of the number N of servers under different service rates μ is plotted in Figure 3. Table 2 contains the values of the objective function.

The optimal values J^* of the objective function for each of the five service rates are displayed in bold face. It is seen from Figure 3 and Table 2 that as the service rate decreases from 5 to 1, the optimal number of servers N^* increases from 3 to 6. The relative gain of the optimal configuration in comparison to a system with an arbitrary number N of the servers at Station 2 is defined as $R_{\text{rel}}(N) = ((J(N) - J^*)/J^*)100\%$.

We now focus on the result of optimal value in the case $\mu = 5$. It is seen from Figure 3 and Table 2 that the optimal value of the objective function J^* is 20.6657 and the optimal number of servers $N^* = 3$. It should be also noted that for the case under consideration the minimal relative gain is more than 10% if we install the optimal number of servers $N^* = 3$ instead of 2 servers and maximal relative gain is more than 161% if we use $N^* = 3$ servers instead of 10 servers.

Experiment 3. In this experiment, we show that the correlation in the input flow has a great impact on the performance measures of the system. In addition to the BMAP defined in the first experiment and having the coefficient of correlation $c_{\text{cor}} = 0.2$, let us consider two another BMAPs, having the same mean arrival rate, but different coefficients of correlation. These

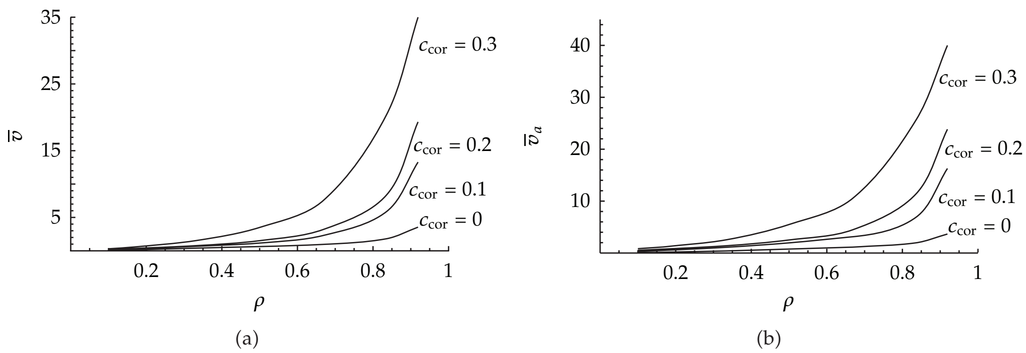

Figure 4: The mean virtual and actual sojourn time as functions of the system load for the BMAPs with different correlation.

BMAPs are defined by the matrices D_0 and $D_1 = D$, from which the matrices $D_k,\ k = \overline{0,5}$, are defined in the same way as in Experiment 1.

The BMAP having the coefficient of correlation $c_{\text{cor}} = 0.1$ is characterized by the matrices

$$D_0 = \begin{pmatrix} -13.3346 & 0.5886 & 0.6173 \\ 0.6927 & -2.4466 & 0.4229 \\ 0.6823 & 0.4144 & -1.6354 \end{pmatrix}, \qquad D = \begin{pmatrix} 11.5469 & 0.3631 & 0.2187 \\ 0.3842 & 0.8659 & 0.0809 \\ 0.2852 & 0.0425 & 0.2111 \end{pmatrix}. \tag{5.4}$$

The BMAP having the coefficient of correlation $c_{\text{cor}} = 0.3$ is defined by the matrices

$$D_0 = \begin{pmatrix} -25.5398 & 0.3933 & 0.3612 \\ 0.1452 & -2.2322 & 0.2000 \\ 0.2960 & 0.3874 & -1.7526 \end{pmatrix}, \qquad D = \begin{pmatrix} 24.2421 & 0.4669 & 0.0763 \\ 0.0341 & 1.6668 & 0.1861 \\ 0.0090 & 0.2555 & 0.8047 \end{pmatrix}. \tag{5.5}$$

All the three defined BMAPs have the coefficient of variation $c_{\text{var}} = 2$.

In addition, we consider the BMAP which is a group Poisson process. It has the same mean arrival rate as three other BMAPs, coefficient of correlation $c_{\text{cor}} = 0$ and the coefficient of variation $c_{\text{var}} = 1$.

The service time at Station 1 has the Erlangian distribution of order 3 with parameter 20. The mean service time $b_1 = 3/20$ and the squared coefficient of variation $c_{\text{var}}^2 = 1/3$.

The number of servers at Station 2 $N = 5$. Service rate μ is equal to 5. Probability γ that a customer will await the release of servers is equal to 0.5. The components of the vector $\mathbf{q}$, which defines the type of the customer, are $q_0 = 0.1,\ q_1 = q_2 = 0.3$, and $q_3 = q_4 = q_5 = 0.1$.

Figures 4 and 5 illustrate the dependence of the mean virtual sojourn time $\bar{v}$, the mean actual sojourn time $\bar{v}_a$, the loss probability P_{loss}, and the mean number N_{busy} of busy servers at Station 2 on the system load ρ. The load ρ varies by means of scaling the fundamental rate λ. Note that the coefficients of correlation and variation of the BMAP do not change under such scaling.

Figures 4 and 5 confirm the fact that values of $\bar{v}$, $\bar{v}_a$, P_{loss}, and N_{busy} increase when the system load, ρ, increases. We also note that, under the same scenario for the system load

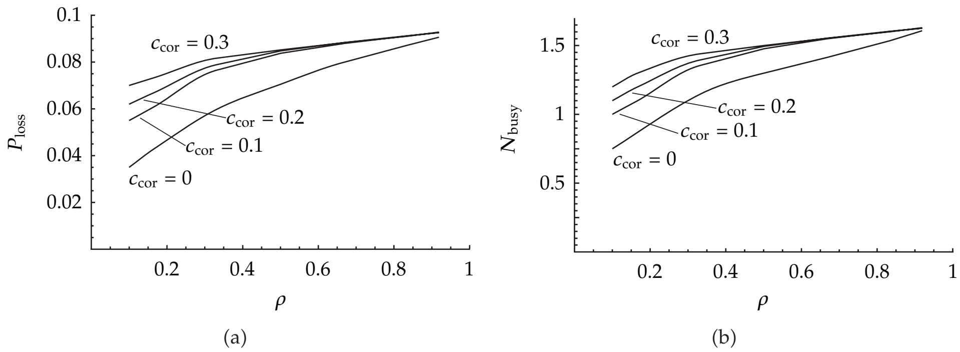

Figure 5: The loss probability and the mean number of busy servers at Station 2 as functions of the system load for the BMAPs with different correlation.

the correlation of the interarrival times shows a strong (negative) impact on the system performance characteristics.

6. Conclusion

In this paper, the BMAP$/G/1 \to \bullet/M/N/0$ tandem queue with heterogeneous customers is studied. The system is studied by looking at selected embedded epochs. The condition for the existence of the stationary distribution is derived. Expressions for the loss probability, blocking probability, and some other performance characteristics of the system are obtained. The Laplace-Stieltjes transforms of the distribution of the virtual and the actual sojourn time at both stations as well as at the whole system are derived. Although the required analytical derivations are very complicated and cumbersome, the resulting formulas have very simple forms. The procedure for calculating the moments of the virtual and the actual sojourn time distribution is elaborated. Illustrative numerical results highlight the important role played by the variance of the service time. An optimization problem to illustrate the usefulness of such problems in practice involving tandem queues is discussed. The results of this paper can be applied to areas such as capacity planning, performance evaluation, and optimization of real-world tandem queues and two-node networks.

Appendices

A. Expressions for the Matrices $\widetilde{K}_j^*(i,1,s,t),\ j>0,\ i>0$

$$\widetilde{K}_0^*(i,1,s,t) = \left(\widetilde{Q}_1 + \gamma\widetilde{Q}_2\right)\int_0^t \left[I_{N+1} \otimes e^{D_0 x}\sum_{k=1}^{i} D_k\, dx P(i-k, t-x)\right]$$

$$\times \int_0^\infty e^{-su} e^{\Delta(t+u)} \otimes I_{\overline{W}}\, dB(t-x+u) + (1-\gamma)$$

$$
\begin{aligned}
&\times \left\{ \int_0^t \left[dF(x) \otimes \sum_{k=1}^{i} P(k,x)P(i-k,t-x) \right] \widetilde{Q}_3 \right.\\
&\qquad \times \int_0^\infty e^{-su} e^{\Delta(t-x+u)} \otimes I_{\overline{W}} dB(t-x+u)\\
&\qquad + \int_0^t \int_0^y \left[dF(x) \otimes \left(e^{D_0 y} \sum_{k=1}^{i} D_k dy P(i-k,t-y) \right) \right] \widetilde{Q}_3\\
&\qquad \left. \times \int_0^\infty e^{-su} e^{\Delta(t-x+u)} \otimes I_{\overline{W}} dB(t-y+u) \right\}, \quad i>0,\\
\widetilde{K}_j^*(i,1,s,t) &= \left(\widetilde{Q}_1 + \gamma \widetilde{Q}_2\right) \left[I_{N+1} \otimes P(i-j,t) \right] \int_0^\infty e^{-su} e^{\Delta(t+u)} \otimes I_{\overline{W}} dB(t+u)\\
&\quad + (1-\gamma) \int_0^t \left[dF(x) \otimes \sum_{k=0}^{i-j} P(k,x)P(i-k-j,t-x) \right] \widetilde{Q}_3\\
&\quad \times \int_0^\infty e^{-su} e^{\Delta(t-x+u)} \otimes I_{\overline{W}} dB(t-x+u).
\end{aligned}
\tag{A.1}
$$

B. Proof of Theorem 4.8.

Proof. Successively differentiating the expression in (4.10), we get

$$
\mathbf{v}_1^{(m)}(0)A(0) = \boldsymbol{\pi}_0 \Phi^{(m)}(0) - \sum_{l=0}^{m-1} \binom{m}{l} \mathbf{v}_1^{(l)}(0) A^{(m-l)}(0), \quad m \geq 0. \tag{B.1}
$$

It follows from (4.11) that $A(0) = \sum_{r=0}^{\infty} \mathcal{B}^r(0) \otimes D_r$, where $\mathcal{B}(0)$ is an irreducible stochastic matrix. This implies that $A(0)$ is an irreducible infinitesimal generator, and hence $A(0)$ is a singular matrix. Thus, it is not possible to develop a recursive scheme for computing the vectors $\mathbf{v}_1^{(m)}(0)$, $m \geq 0$, directly from (B.1). We will now modify the system (B.1) to get the system with a nonsingular matrix. To this end, we postmultiply the expression for $m+1$ in (B.1) on both sides with $\mathbf{e}$. Taking into account that $A(0)\mathbf{e} = \mathbf{0}^T$, we get

$$
\mathbf{v}_1^{(m)} A'(0)\mathbf{e} = \frac{1}{m+1} \left[\boldsymbol{\pi}_0 \Phi^{(m+1)}(0) - \sum_{l=0}^{m-1} \binom{m+1}{l} \mathbf{v}_1^{(l)}(0) A^{(m+1-l)}(0) \right] \mathbf{e}. \tag{B.2}
$$

It can be shown that the right-hand side of (B.2) is not equal to zero. It is positive if $m = 2k$ and it is negative if $m = 2k+1$, $k \geq 0$. Thus, replacing one of the equations in the system (B.1) (without loss of generality, we replace the first equation) with equation (B.2), we get the following (inhomogeneous) system of linear algebraic equations for the entries of the vector $\mathbf{v}_1^{(m)}(0)$:

$$\begin{aligned}\mathbf{v}_1^{(m)}(0)\widetilde{A} = {} & \left[\boldsymbol{\pi}_0\Phi^{(m)}(0) - \sum_{l=0}^{m-1}\binom{m}{l}\mathbf{v}_1^{(l)}(0)A^{(m-l)}(0)\right]\widetilde{I} \\ & + \frac{1}{m+1}\left[\boldsymbol{\pi}_0\Phi^{(m+1)}(0) - \sum_{l=0}^{m-1}\binom{m+1}{l}\mathbf{v}_1^{(l)}(0)A^{(m+1-l)}(0)\right]\mathbf{e}\widehat{\mathbf{e}}, \quad m \geq 0.\end{aligned} \tag{B.3}$$

The above system has the unique solution if $\widetilde{A}$ is non-singular. We prove this by showing that $\det\widetilde{A} \neq 0$.

Let us calculate $\det\widetilde{A}$ as

$$\det\widetilde{A} = \nabla A'(0)\mathbf{e}, \tag{B.4}$$

where ∇ is a vector of algebraic cofactors of the first column of the matrix $A(0)$. Since $A(0)$ is irreducible, the vector ∇ is proportional to any solution of the system

$$\mathbf{x}A(0) = \mathbf{0}, \tag{B.5}$$

that is,

$$\nabla = c\mathbf{x}, \tag{B.6}$$

where the scalar $c \neq 0$.

Let the vector $\boldsymbol{\vartheta}$ be the unique solution to the system

$$\boldsymbol{\vartheta}\mathcal{B}(0) = \vartheta, \qquad \boldsymbol{\vartheta}\mathbf{e} = 1. \tag{B.7}$$

Then, by the direct substitution it can be verified that the vector $\mathbf{x} = \vartheta \otimes \boldsymbol{\theta}$ is a solution to the system (B.5).

From (B.6), $\nabla = c(\boldsymbol{\vartheta} \otimes \boldsymbol{\theta})$. Substituting the vector ∇ into (B.4), we obtain

$$\begin{aligned}\det\widetilde{A} &= c(\boldsymbol{\vartheta}\otimes\boldsymbol{\theta})A'(0)\mathbf{e} = c(\boldsymbol{\vartheta}\otimes\boldsymbol{\theta})\left(sI + \sum_{r=0}^{\infty}\mathcal{B}^r(s)\otimes D_r\right)'\Bigg|_{s=0}\mathbf{e} \\ &= c + c\sum_{r=1}^{\infty} r\boldsymbol{\vartheta}\mathcal{B}'(0)\mathbf{e}\otimes\boldsymbol{\theta}D_r\mathbf{e} = c\left(1 + \lambda\boldsymbol{\vartheta}\mathcal{B}'(0)\mathbf{e}\right).\end{aligned} \tag{B.8}$$

In further evaluation of $\det\widetilde{A}$, we use the ergodicity condition given in Theorem 3.3.

Setting $s = 0$ in (4.9) and noting that $Q_1 + \gamma Q_2 + (1-\gamma)F^*(0)Q_3 = Q$, we see that $\mathcal{B}(0) = QB^*(0)$ and that the vector ϑ defined by (B.7) is the unique solution of system (3.8). It can be easily verified that

$$\boldsymbol{\vartheta}\mathcal{B}'(0)\mathbf{e} = -\left[b_1 + (1-\gamma)\boldsymbol{\vartheta}\int_0^{\infty} t dF(t) Q_3 \mathbf{e}\right]. \tag{B.9}$$

Multiplying (B.9) by λ and comparing the obtained equation with the equality in (3.12) we see that

$$\lambda\boldsymbol{\vartheta}\mathcal{B}'(0)\mathbf{e} = -\rho. \tag{B.10}$$

It follows from (B.8) and (B.10) that $\det \widetilde{A} = c(1-\rho)$. Since the stability condition implies that $\rho < 1$ and since $c \neq 0$, we have $\det \widetilde{A} \neq 0$. This competes the proof of the theorem.

C. Formulas for Calculation of the Mean Virtual Sojourn Time

The expressions (4.21)-(4.22) for the mean virtual sojourn time involve the matrices $\widetilde{A}$, $A'(0)$ and the vector $A''(0)\mathbf{e}$ for which we derive explicit expressions below.

$$\begin{gathered}
\widetilde{A} = \left[\sum_{r=0}^{\infty} \mathcal{B}^r(0) \otimes D_r\right]\widetilde{I} + \left[I + \sum_{r=1}^{\infty}\sum_{n=0}^{r-1} \mathcal{B}^n(0)\mathcal{B}'(0) \otimes D_r\right]\mathbf{e}\widehat{\mathbf{e}}, \\
A'(0) = I + \sum_{r=1}^{\infty}\sum_{n=0}^{r-1}\left(\mathcal{B}^n(0)\mathcal{B}'(0)\mathcal{B}^{r-n-1}(0) \otimes D_r\right), \\
A''(0)\mathbf{e} = \left\{\sum_{r=1}^{\infty}\sum_{n=1}^{r-1}\left[2\sum_{l=0}^{n-1}\mathcal{B}^l(0)\mathcal{B}'(0)\mathcal{B}^{n-l-1}(0)\mathcal{B}'(0) + \mathcal{B}^n(0)\mathcal{B}''(0)\right] \otimes D_r\right\}\mathbf{e},
\end{gathered} \tag{C.1}$$

where

$$\begin{gathered}
\mathcal{B}(0) = QB^{*(0)}, \qquad \mathcal{B}'(0) = QB^{*(1)} + (1-\gamma)F^{(1)}Q_3B^{*(0)}, \\
\mathcal{B}''(0) = QB^{*(2)} - 2(1-\gamma)F^{(1)}Q_3B^{*(1)} + (1-\gamma)F^{(2)}Q_3B^{*(0)},
\end{gathered} \tag{C.2}$$

$$F^{(m)} = (-1)^m \int_0^{\infty} t^m dF(t) = \left(f_{r,r'}^{(m)}\right)_{r,r'=\overline{0,N}}, \quad m = 1,2, \tag{C.3}$$

$$f_{r,r'}^{(1)} = \begin{cases} 0, & r \le r', \\ -\sum_{l=r'+1}^{r} (l\mu)^{-1}, & r > r', \end{cases} \tag{C.4}$$

$$f^{(2)}_{r,r'} = \begin{cases} 0, & r \le r', \\ 2\displaystyle\sum_{l=r'+1}^{r} (-1)^{l-r'+1} \binom{r-r'}{l-r'} (l\mu)^{-2}, & r > r', \end{cases}$$

$$B^{*(m)} = (-1)^m \int_0^\infty t^m e^{\Delta t} dB(t) = \left(\beta^{*(m)}_{r,r'}\right)_{r,r'=\overline{0,N}}, \qquad m = \overline{0,2}, \tag{C.5}$$

$$\beta^{*(m)}_{r,r'} = \begin{cases} 0, & r < r', \\ \displaystyle\binom{r}{r'} \left\{ \sum_{i=0}^{r-r'} (-1)^i \binom{r-r'}{i} \beta^{(m)}\left(\mu(r'+i)\right) \right\}, & r \ge r'. \end{cases}$$

Remark C.1. Formulas for $\widetilde{A}, A'(0), A''(0)$ contain infinite sums. However, the calculations of these should not create any difficulty as for overwhelming majority of interesting and useful queueing models; the parameter matrices, D_k, of the BMAP process are equal to zero for k greater than some threshold, say, K. Thus, all sums become finite. In alternative case, some analytical formula for computing the infinite sequence of matrices $D_k,\ k \ge 1$, should be given. If this sequence is generated such as $D_k = D(1-\sigma)\sigma^{k-1},\ k \ge 1$, where D and $0 < \sigma < 1$ are given parameters, the infinite sums can be computed explicitly.

Acknowledgments

The authors are thankful to Professor S. R. Chakravarthy (Kettering University, USA) for valuable remarks and suggestions and kind help in polishing the text of this paper. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant no. 2010-0003269).

References

[1] S. Balsamo, V. De Nitto Personè, and P. Inverardi, "A review on queueing network models with finite capacity queues for software architectures performance prediction," *Evaluation*, vol. 51, no. 2–4, pp. 269–288, 2003.

[2] H. W. Ferng and J. F. Chang, "Connection-wise end-to-end performance analysis of queueing networks with MMPP inputs," *Performance Evaluation*, vol. 43, no. 1, pp. 39–62, 2001.

[3] A. Heindl, "Decomposition of general tandem queueing networks with MMPP input," *Performance Evaluation*, vol. 44, no. 1–4, pp. 5–23, 2001.

[4] B. W. Gnedenko and D. König, *Handbuch der Bedienungstheorie*, vol. 56, Akademie-Verlag, Berlin, Germany, 1983.

[5] D. M. Lucantoni, "New results on the single server queue with a batch Markovian arrival process," *Communications in Statistics. Stochastic Models*, vol. 7, no. 1, pp. 1–46, 1991.

[6] A. Gómez-Corral, "A tandem queue with blocking and Markovian arrival process," *Queueing Systems*, vol. 41, no. 4, pp. 343–370, 2002.

[7] A. Gómez-Corral, "On a tandem G-network with blocking," *Advances in Applied Probability*, vol. 34, no. 3, pp. 626–661, 2002.

[8] V. Klimenok, L. Breuer, G. Tsarenkov, and A. Dudin, "The $BMAP/G/1/N \to \bullet/PH/1/M$ tandem queue with losses," *Performance Evaluation*, vol. 61, no. 1, pp. 17–40, 2005.

[9] C. S. Kim, V. Klimenok, G. Tsarenkov, L. Breuer, and A. Dudin, "The $BMAP/G/1 \to \bullet/PH/1/M$ tandem queue with feedback and losses," *Performance Evaluation*, vol. 64, no. 7-8, pp. 802–818, 2007.

[10] A. Gómez-Corral and M. E. Martos, "Performance of two-stage tandem queues with blocking: the impact of several flows of signals," *Performance Evaluation*, vol. 63, no. 9-10, pp. 910–938, 2006.
[11] D. Menasce and V. Almeida, *Capacity Planning for Web Performance: Metrics, Models, and Methods*, Prentice Hall, New York, NY, USA, 1998.
[12] T. Janevski, *Traffic Analysis and Design of Wireless IP Networks*, Artech House, Boston, Mass, USA, 2003.
[13] A. Gómez-Corral, "Sojourn times in a two-stage queueing network with blocking," *Naval Research Logistics*, vol. 51, no. 8, pp. 1068–1089, 2004.
[14] A. Gómez-Corral and M. E. Martos, "A. B. Clarke's tandem queue revisited—sojourn times," *Stochastic Analysis and Applications*, vol. 26, no. 6, pp. 1111–1135, 2008.
[15] S. R. Chakravarthy, "The batch Markovian arrival process: a review and future work," in *Proceedings of the International Conference on Advances in Probability Theory and Stochastic Processes*, A. Krishnamoorthy et al., Ed., pp. 21–49, Notable Publications, 2001.
[16] M. F. Neuts, *Structured Stochastic Matrices of* $M/G/1$ *Type and Their Applications*, vol. 5, Marcel Dekker, New York, NY, USA, 1989.
[17] E. Çinlar, *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1975.