J. Korean Soc. Ind. Appl. Math. Vol.26, No.1, 1–22, 2022 https://doi.org/10.12941/jksiam.2022.26.001

AN APPROXIMATE ANALYSIS OF TANDEM QUEUES WITH GENERAL BLOCKING NODES

YANG WOO SHIN[1†], DONG OK KIM[2], AND DUG HEE MOON[3]

[1]DEPARTMENT OF STATISTICS, CHANGWON NATIONAL UNIVERSITY, CHANGWON, 51140, REPUBLIC OF KOREA
Email address: [†]ywshin@changwon.ac.kr

[2]INSTITUTE OF INDUSTRIAL TECHNOLOGY, CHANGWON NATIONAL UNIVERSITY, CHANGWON, 51140, REPUBLIC OF KOREA

[3]DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING, CHANGWON NATIONAL UNIVERSITY, CHANGWON, 51140, REPUBLIC OF KOREA

ABSTRACT. A tandem queue that consists of nodes with buffers of finite capacity and general blocking scheme is considered. The service time distribution of each node is exponential whose rate depends on the state of the node. The blocking scheme at a node may be different from that of other nodes. An approximation method for the system based on decomposition method is presented. The effectiveness of the method is investigated numerically.

1. INTRODUCTION

Tandem queue is a queueing network in which service nodes are linked along a single flow path one after another and customers arrive from outside at the first node and are processed at the nodes in sequence, and leave the system from the last node. Tandem queues with finite capacity nodes have been widely used for performance modeling of computer systems, telecommunication networks and manufacturing systems [1, 2, 3]. The limited buffer capacity leads to the blocking phenomenon. When a node reaches its maximum capacity, the flow of customers from the upstream node into the downstream node is stopped, and the blocking phenomenon arises. Various blocking mechanisms in queueing networks with finite capacity nodes have been introduced in the literature to represent distinct behaviors of real systems [4]. The blocking type mostly used in modelling manufacturing systems is *blocking after service* (BAS) scheme, or sometimes called *manufacturing blocking*, in which if the buffer of the destination node is full upon a service completion at a node, the server is forced to stop its service, and the customer is held at the node where it has recently completed its service until the destination

Received November 25 2021; Revised March 7 2022; Accepted in revised form March 8 2022; Published online March 25 2022.

2010 *Mathematics Subject Classification.* 60K25, 68M20.

Key words and phrases. tandem queue, general blocking, decomposition.

[†] Corresponding author. This paper was supported by Changwon National University in 2021-2022.

can accommodate it. Under the *blocking before service* (BBS) scheme, sometimes called *communication blocking*, the server at each node checks the state of the destination node before starting a service and if there is an available space at the destination node, then the server starts its service, otherwise, the server is blocked and does not start its service. Two subcategories of BBS scheme distinguish whether the server can be used as a buffer when the node is blocked, called BBS-SO (server occupied) or not, called BBS-SNO (server not occupied). Another type of blocking scheme is *kanban* blocking under which the customers blocked upon a service completion share the buffer space of the node along with the other customers that are either waiting for service or being served, and the server continues processing customers in the node unless the server is not blocked. Cheng and Yao [5] develop a *general blocking* (GB) scheme by introducing parameters for the upper limits on the number of customers waiting in buffer and being in service, the number of blocked customers, and the capacity of the node. The GB scheme includes ordinary manufacturing, communication, and kanban blocking as special cases by specifying the values of the parameters in different ways. The structural properties in tandem queues with GB scheme such as the line reversibility and the effects of system parameters, buffer size, variability of service time and the control parameters of the blocking scheme to the system performance are presented in [5, 6, 7]. In this paper, we focus on the quantitative (numerical) method for the analysis of the tandem queues with exponential service time under GB scheme.

Models with finite buffers and exponential service times can be represented by finite state Markov chains. However, a numerical solution of the associated Markov chain is seriously limited by the complexity of state space and computational time that grow exponentially as the number of nodes increases. Hence, approximate analytical methods and simulation have been used for numerical analysis of the system as alternatives of exact solution.

Many approximation methods for queueing networks with blocking have been proposed in literature both for open and closed models and surveys of some methods have been presented [8, 9, 1, 2]. One of the most common method among the approximation techniques is decomposition method in which the original long line is decomposed into subsystem that are mathematically tractable, and the performance of original system is approximated by that of the subsystems. For more about decomposition method for tandem queues, see e.g. [10, 11, 12, 13]. Most of the works cited above pertain queueing systems under BAS blocking strategy.

The objective of this paper is to present an approximate analysis for the tandem queue with finite buffer under general blocking scheme. The processing time at each node is exponentially distributed and the service rate depends both on the number of customers that are waiting for service in queue or being served and the number of blocked customers. This system includes the tandem queue with multiple servers and general blocking at each node as a special case by specifying the service rates. Our approach is based on decomposition method. The contribution of this paper is to present an approximation method for a very general model in blocking mechanism sense. The method is very effective in accuracy and computation time.

The paper is organized as follows. In Section 2 we describe the model in detail. Some preliminaries and subsystem are presented in Sections 3 and 4, respectively. Approximate formulae for the parameters of subsystems are presented in Section 5 and an algorithm for

calculating performance measures is given in Section 6. Application of the result to the system with multiple servers is described in Section 7. The results of the approximation method are compared numerically with simulation and existing methods in Section 8. Finally, some concluding remarks are given in Section 9.

2. MODEL AND ASSUMPTIONS

We consider a tandem queueing network L that consists of N nodes W_i with finite capacity $c_i < \infty$, $i = 1, 2, \cdots, N$. The customers arrive from outside according to a Poisson process whose rate depends on the state of the first node and customers at W_i are processed (served) according to an exponential distribution whose rate depends on the state of W_i. The blocked customer is the one that has completed service at a node, but cannot be sent to the next node due to a limited buffer capacity of the downstream node. The blocked customers may continue to share the buffer space of the node along with the other customers that are either waiting for service or being served upon. We classify the customers at node W_i into two types, blocked customers (BC) and active customers (AC) that are waiting for service or being in service at node W_i.

Blocking scheme. The blocking process of each node is controlled by three parameters (a_i, b_i, c_i), where a_i and b_i are the upper limits on the number of active customers and blocked customers at W_i, respectively, and c_i is the capacity of W_i. It is natural to assume that

$$1 \le a_i \le c_i, \;\; 0 \le b_i \le c_i, \;\; a_i + b_i \ge c_i.$$

We assume that the last node is never blocked and $b_N = 0$.

The features of the node with $b_i > 0$ are different from those of the node with $b_i = 0$. In case of $b_i = 0$, W_i cannot hold any blocked customers and the node follows the BBS rule. If $b_i > 0$, then the behavior (blocking or joining to the next node) of the customers in W_i is determined after a service completion. We describe the behaviors of two types of nodes separately.

(i) The case of $b_i > 0$. Upon a service completion at W_i, if there are no places available for active customers in W_{i+1}, that is, the number of active customers at W_{i+1} is a_{i+1} or total number of customers at W_{i+1} is c_{i+1}, then the customer just completed its service is blocked and is stocked at W_i. If the number of blocked customers at W_i reaches b_i upon a service completion, then the service process at W_i is forced to stop. Even the service process is stopped, the node W_i can hold active customers arriving from the upstream node W_{i-1} if there is an available space for active customers at W_i.

(ii) The case of $b_i = 0$. In this case, if the number of active customers in W_{i+1} reaches a_{i+1} or total number of customers in W_{i+1} reaches c_{i+1}, the service process at W_i is forced to stop until there is an available space for active customers in W_{i+1}. The idle server can accept an active customer even it is stopped its service, that is, W_i follows BBS-SO blocking scheme. The node W_i starts new service upon there is a place available for active customers at W_{i+1}. Hereafter BBS means BBS-SO blocking. Define the state of service process at W_i at time t by

$$M_i(t) = \begin{cases} 0^*, & \text{service process at } W_i \text{ is stopped,} \\ 0, & \text{otherwise.} \end{cases}$$

Service time. The distribution of service time at node W_i is exponential and the service rate may depend on the state of the node W_i. Denote the service rate by $\mu_i(x, y)$, where x is the number of AC and y is the number of BC for $b_i > 0$, and for $b_i = 0$, y is the server state at W_i. It can be seen from the assumption that $\mu_i(x, b_i) = 0$ for $b_i > 0$ and $\mu_i(x, 0^*) = 0$ for $b_i = 0$, and $\mu_i(0, y) = 0$.

Arrival process from source node. Customers arrive to the first node W_1 from the source node (or outside) W_0 according to a Poisson process whose rate depends on the state of W_1. The node W_0 is assumed to be never starved. Denote the arrival rate to the first node W_1 by $\lambda_1(x, y)$, where x is the number of AC, and for $b_1 > 0$, y is the number of BC at W_1 and for $b_1 = 0$, $y \in \{0, 0^*\}$ is the server state of W_1.

3. Stochastic processes and transition rates

3.1. **Stochastic processes.** Let $X_i^a(t)$ and $Y_i(t)$ be the number of active customers and blocked customers, respectively, in W_i at time t, and $X_i(t) = X_i^a(t) + Y_{i-1}(t)$, $Z_i(t) = X_i(t) + Y_i(t)$.

For describing the behavior of the node W_i, we use two dimensional stochastic processes

$$\begin{aligned} V_i(t) &= (Z_i(t), X_i(t)), \\ W_i(t) &= \begin{cases} (X_i(t), Y_i(t)), & b_i > 0, \\ (X_i(t), M_i(t)), & b_i = 0. \end{cases} \end{aligned}$$

Let $\xi_i = a_i + b_{i-1}$ and $\kappa_i = c_i + b_{i-1}$, $i = 1, 2, \cdots, N$. The state space of $Z_i(t)$ is $\mathcal{Z}_i = \{0, 1, \cdots, \kappa_i\}$. Once $Z_i(t) = n$ is given, it can be seen from $0 \le X_i(t) \le \xi_i$ and $0 \le Y_i(t) \le b_i$ that $l_i(n) \le X_i(t) \le u_i(n)$, where

$$l_i(n) = \max(n - b_i, 0), \quad u_i(n) = \min(n, \xi_i).$$

Note that if $Y_i(t) = y$ is given, then $0 \le X_i(t) \le x_i^*(y)$, where

$$x_i^*(y) = \min(\xi_i, \kappa_i - y).$$

The state space $\mathcal{V}_i$ of $V_i(t)$ and the space $\mathcal{W}_i$ of $W_i(t)$ are as follow:

$$\begin{aligned} \mathcal{V}_i &= \{(n, x) : l_i(n) \le x \le u_i(n), 0 \le n \le \kappa_i\}, \\ \mathcal{W}_i &= \begin{cases} \{(x, y) : 0 \le x \le x_i^*(y),\ 0 \le y \le b_i\}, & b_i > 0, \\ \{(x, 0)\,(x, 0^*) : 0 \le x \le \xi_i\}, & b_i = 0. \end{cases} \end{aligned}$$

Note that for given $V_i(t) = (n, x)$, the maximal value $x_{i-1}(n, x)$ of $X_{i-1}(t)$ and the state $y_{i-1}(n, x)$ of $Y_{i-1}(t)$ for $b_{i-1} > 0$ are determined by

$$\begin{aligned} y_{i-1}(n, x) &= \max(0, n - c_i, x - a_i), \\ x_{i-1}(n, x) &= \max(\xi_{i-1}, \kappa_{i-1} - y_{i-1}(n, x)). \end{aligned}$$

and the state of $M_{i-1}(t)$ for $b_{i-1} = 0$ is

$$M_{i-1}(t) = \begin{cases} 0, & n < c_i \text{ and } x < a_i, \\ 0^*, & n = c_i \text{ or } x = a_i. \end{cases}$$

Let

$$\mathcal{D}_{i-1} = \{(n,x) \in \mathcal{V}_i \,:\, l_i(n) \le x \le \min(n, a_i - 1), 0 \le n \le c_i - 1\}$$

and for $0 \le y \le b_{i-1}$,

$$\begin{aligned}\mathcal{B}_{i-1}(y) &= \{(c_i + y, x) \in \mathcal{V}_i \,:\, l_i(c_i + y) \le x \le a_i + y\} \\ &\quad \cup \{(n, a_i + y) \in \mathcal{V}_i \,:\, a_i + y \le n < c_i + y\}.\end{aligned}$$

Noting that $\mathcal{D}_{i-1}$ is the set of states of $V_i(t)$ on which there is an available space for active customers in W_i and $\mathcal{B}_{i-1}(0)$ is the set on which there are no places available for active customers in W_i and $Y_{i-1}(t) = 0$, it can be seen that for $b_{i-1} > 0$,

$$\begin{aligned}\{Y_{i-1}(t) = 0\} &= \{V_i(t) \in \mathcal{D}_{i-1} \cup \mathcal{B}_{i-1}(0)\}, \\ \{Y_{i-1}(t) = y\} &= \{V_i(t) \in \mathcal{B}_{i-1}(y)\},\ 1 \le y \le b_{i-1},\end{aligned} \tag{3.1}$$

and for $b_{i-1} = 0$,

$$\begin{aligned}\{M_{i-1}(t) = 0\} &= \{V_i(t) \in \mathcal{D}_{i-1}\}, \\ \{M_{i-1}(t) = 0^*\} &= \begin{cases} \{V_i(t) \in \mathcal{B}_{i-1}(0)\}, & b_i > 0, \\ \{W_i(t) \in \{(c_i, 0), (c_i, 0^*)\}\}, & b_i = 0. \end{cases}\end{aligned} \tag{3.2}$$

We introduce notation for later use. For given $V_i(t) = (n, x)$, we write the state of $Y_{i-1}(t)$ for $b_{i-1} > 0$ and $M_{i-1}(t)$ for $b_{i-1} = 0$ by a unified form

$$y^*_{i-1}(n,x) = \begin{cases} y_{i-1}(n,x), & b_{i-1} > 0, \\ 0, & (n,x) \in \mathcal{D}_{i-1}, b_{i-1} = 0, \\ 0^*, & (n,x) \in \mathcal{B}_{i-1}(0), b_{i-1} = 0 \end{cases}$$

and $y^*_{i-1}(x + 0^*, x)$ means $y^*_{i-1}(x, x)$.

3.2. **Transition rates of $W_i(t)$.** Given $W_i(t) = (x, y) \in \mathcal{W}_i$, the state transitions of $W_i(t)$ are occurred by a service completion at W_{i-1} or W_i, a departure of blocked customers from W_i for $b_i > 0$, and a departure from W_{i+1} for $b_i = 0$ and $y = 0^*$. Now we derive the transition rates of $W_i(t)$ for each case described above.

(i) The rate $\lambda_i(x, y)$ from (x, y) to $(x + 1, y)$. The transition from (x, y) to $(x + 1, y)$ is occurred by a service completion at W_{i-1} and the rate is for $2 \le i \le N - 1$,

$$\lambda_i(x,y) = \sum_{j=1}^{h} P(X_{i-1}(t) = j | W_i(t) = (x,y)) \mu_{i-1}(j,k),$$

where $h = x_{i-1}(x + y, x)$ and $k = y^*_{i-1}(x + y, x)$.

(ii) The rate $\beta_i(x, y)$ from (x, y) to $(x - 1, y + 1)$ for $b_i > 0$. Given $W_i(t) = (x, 0)$, the customer being served at W_i is blocked to enter the next node W_{i+1} upon a service completion if there are no places available for active customers. Thus for $1 \le x \le \xi_i$,

$$\beta_i(x,0) = P(V_{i+1}(t) \in \mathcal{B}_i(0) | W_i(t) = (x,0)) \mu_i(x,0).$$

If $1 \le y \le b_i - 1$, the customer is blocked upon it service completion and hence

$$\beta_i(x,y) = \mu_i(x,y),\ 1 \le x \le x^*_i(y),\ 1 \le y \le b_i - 1.$$

(iii) The rate $\delta_i(x,0)$ from $W_i(t)=(x,0)$ to $(x-1,0)$ for $b_i>0$. It can be seen from (3.1) that

$$\begin{aligned}\delta_i(x,0) &= P(V_{i+1}(t)\in\mathcal{D}_i|W_i(t)=(x,0))\mu_i(x,0)\\ &= \mu_i(x,0)-\beta_i(x,0),\ 1\le x\le\xi_i.\end{aligned}$$

(iv) The rate $\beta_i^0(x)$ from working state $(x,0)$ to blocking state $(x-1,0^*)$ in the node with $b_i=0$. Let

$$\begin{aligned}\mathcal{D}_i^0 &= \{(c_{i+1}-1,x)\in\mathcal{V}_{i+1}\,:\,l_{i+1}(c_{i+1}-1)\le x\le a_{i+1}-1\}\\ &\quad\cup\{(n,a_{i+1}-1)\,:\,a_{i+1}-1\le n<c_{i+1}-1\}.\end{aligned}$$

If $V_{i+1}(t)\in\mathcal{D}_i^0$, then there is only one place available for active customers in W_{i+1} at time t. The transition of $W_i(t)$ from $(x,0)$ to $(x-1,0^*)$ occurs if a service at W_i is completed on $\mathcal{D}_i^0$. Thus the rate from $(x,0)$ to $(x-1,0^*)$ is for $1\le x\le\xi_i$,

$$\beta_i^0(x)=P\left(V_{i+1}(t)\in\mathcal{D}_i^0|W_i(t)=(x,0)\right)\mu_i(x,0)$$

(v) The rate $\delta_i^0(x)$ from $(x,0)$ to $(x-1,0)$ in the node with $b_i=0$. In case of $b_i=0$, customers at W_i join W_{i+1} upon a service completion and the resulting state of $M_i(t)$ is one of the two states 0 or 0^*. It can be seen from (3.2) that

$$\delta_i^0(x)=\mu_i(x,0)-\beta_i^0(x),\ 1\le x\le\xi_i.$$

(vi) The rate $\delta_i(x,y)$ from $W_i(t)=(x,y)$ to $(x,y-1)$, $y\ge1$. Let

$$\begin{aligned}\tilde{\mathcal{B}}_i(y) &= \{(n',x')\in\mathcal{B}_i(y)\,:\,y_i(n'-1,x')=y-1,n'>x'\}\\ &\quad\cup\{(a_{i+1}+y,a_{i+1}+y)\}.\end{aligned}$$

If $V_{i+1}(t)\in\tilde{\mathcal{B}}_i(y)$ with $y\ge1$ and a departure from the node W_{i+1} occurs, then the resulting state of $V_{i+1}(t)$ is in $\tilde{\mathcal{B}}_i(y-1)$. If $W_i(t)=(x,y)$ with $y\ge1$, a departure from W_i is occurred by a departure from W_{i+1} on $\tilde{\mathcal{B}}_i(y)$ and hence for $y\ge1$

$$\delta_i(x,y)=\sum_{(n',x')\in\tilde{\mathcal{B}}_i(y)}P(V_{i+1}(t)=(n',x')|W_i(t)=(x,y))\delta_{i+1}^*(x',n'-x'),$$

where

$$\delta_{i+1}^*(x',n'-x')=\begin{cases}\delta_{i+1}(x',n'-x'), & b_{i+1}>0,\\ \delta_{i+1}^0(x'), & b_{i+1}=0\end{cases}$$

with $\delta_N(x,0)=\mu_N(x,0)$ and $\delta_N(x,y)=0$, $y\ge1$. Noting that for $b_{i+1}=0$, $\mathcal{B}_i(y)=\{(c_{i+1}+y,c_{i+1}+y)\}$, it can be seen that for $b_{i+1}=0$,

$$\delta_i(x,y)=\mu_{i+1}(c_{i+1}+y,0),\ 1\le y\le b_i.$$

(vii) The rate $\alpha_i(x)$ from blocking state $(x,0^*)$ to working state $(x,0)$ in the node with $b_i=0$. The transition of $W_i(t)$ from $(x,0^*)$ to $(x,0)$ is occurred by one of two types of departures from W_{i+1}, a departure of an active customer on $V_{i+1}(t)=(a_{i+1},a_{i+1})$ and a

departure of a blocked customer on $V_{i+1}(t) = (c_{i+1}, x')$ with $x' < a_{i+1}$. It can be seen that for $b_{i+1} > 0$,

$$\begin{aligned}\alpha_i(n) &= P(V_{i+1}(t) = (a_{i+1}, a_{i+1}) | W_i(t) = (n, 0^*))\delta_{i+1}(a_{i+1}, a_{i+1}) \\ &+ \sum_{x=l_{i+1}(c_{i+1})}^{a_{i+1}-1} P(V_{i+1}(t) = (c_{i+1}, x) | W_i(t) = (n, 0^*)\delta_{i+1}(x, c_{i+1} - x)\end{aligned}$$

and for $b_{i+1} = 0$,

$$\alpha_i(n) = \mu_{i+1}(c_{i+1}, 0).$$

4. SUBSYSTEMS

We decompose the system into $N-1$ subsystems L_i, $i = 2, 3, \cdots, N$ for an approximate analysis using decomposition method. The subsystem L_i is a tandem queue that consists of two nodes, say W_{i-1}^u and W_i^d which are pseudo nodes that correspond to W_{i-1} and W_i, respectively. The node W_{i-1}^u has a buffer of size $g_{i-1}^u = g_{i-1} + b_{i-2}$, and the capacity of W_{i-1}^u is $\kappa_{i-1}^u = \kappa_{i-1}$, the capacity for active customers is $a_{i-1}^u = \xi_{i-1}$, the capacity for blocked customers is $b_{i-1}^u = b_{i-1}$. Customers arrive to the first node W_{i-1}^u according to a Poisson process with rate $\lambda_{i-1}(j,k)$ and the service rate is $\mu_{i-1}(j,k)$, when $W_{i-1}(t) = (j,k)$. The parameters for W_i^d are the same as W_i. The throughput of L is approximated with that of L_N. Since W_1 and W_N have different features from that of W_i, $2 \le i \le N-1$, we describe the subsystems L_i $(2 \le i \le N-1)$ and L_N separately.

4.1. The subsystem L_i with $b_i > 0$, $2 \le i \le N-1$.

For describing the subsystem L_i with $b_i > 0$, define the stochastic processes $\Psi_i(t) = (Z_i(t), X_i(t), X_{i-1}(t))$, $2 \le i \le N-1$. The state space $\mathcal{S}_i$ of $\mathbf{\Psi}_i = \{\Psi_i(t), t \ge 0\}$ is $\mathcal{S}_i = \cup_{n=0}^{\kappa_i} \mathcal{S}_i(n)$, where

$$\mathcal{S}_i(n) = \{(n, x, j) \,:\, 0 \le j \le x_{i-1}(n, x), l_i(n) \le x \le u_i(n)\},\ 2 \le i \le N-1,$$

It can be easily seen that the number of states in $\mathcal{S}_i(n)$ is

$$s_i(n) = \sum_{x=l_i(n)}^{u_i(n)} (x_{i-1}(n, x) + 1),\ 2 \le i \le N-1.$$

The stochastic process $\mathbf{\Psi}_i = \{\Psi_i(t), t \ge 0\}$ forms a Markov chain on the state space $\mathcal{S}_i$ with generator of the form

$$Q_i = \begin{pmatrix} B_i^{(0)} & A_i^{(0)} & & & \\ C_i^{(1)} & B_i^{(1)} & A_i^{(1)} & & \\ & \ddots & \ddots & \ddots & \\ & & C_i^{(\kappa_i-1)} & B_i^{(\kappa_i-1)} & A_i^{(\kappa_i-1)} \\ & & & C_i^{(\kappa_i)} & B_i^{(\kappa_i)} \end{pmatrix} - \Delta_i, \tag{4.1}$$

where Δ_i is the diagonal matrix that makes $Q_i \mathbf{e} = 0$ and $\mathbf{e}$ is the column vector of appropriate size whose components are all 1. The block matrix $B_i^{(n)}$ is the square matrix of size $s_i(n)$

whose diagonals entries are all 0 and their off-diagonal components correspond to the transition rates in $\mathcal{S}_i(n)$ without changing the level $Z_i(t) = n$. The components of the block matrices $A_i^{(n)}$ and $C_i^{(n)}$ are the transition rates of $\mathbf{\Psi}_i$ from the states of $\mathcal{S}_i(n)$ to the states of $\mathcal{S}_i(n+1)$ and $\mathcal{S}_i(n-1)$, respectively.

In this subsection, for given $(Z_i(t), X_i(t)) = (n, x)$, denote the maximal number of active customers at W_{i-1} by $h = x_{i-1}(n, x)$ for the simplicity of notation unless confusion.

The matrices $A_i^{(n)}$. For each $(n, x), (n+1, x') \in \mathcal{V}_i$, denote by $A_i^{(n)}[x, x']$ the block matrix component of size $(h+1) \times (x_{i-1}(n+1, x') + 1)$ whose (j, j')-entry corresponds to the transition rate from $(n, x, j) \in \mathcal{S}_i(n)$ to $(n+1, x', j') \in \mathcal{S}_i(n+1)$. It can be easily seen that $A_i^{(n)}[x, x'] = 0$ for $x' \neq x+1$ and that for $l_i(n) \leq x \leq u_i(n)$, $l_i(n+1) \leq x+1 \leq u_i(n+1)$, the (j, j')-component of $A_i^{(n)}[x, x+1]$, $0 \leq j \leq h$, $0 \leq j' \leq x_{i-1}(n+1, x+1)$ is

$$\left[A_i^{(n)}[x, x+1]\right]_{jj'} = \mu_{i-1}(j, y_{i-1}^*(n, x))\mathbf{1}(j' = j-1),$$

where $\mathbf{1}(A) = 1$ if A is true, otherwise 0.

The matrices $B_i^{(n)}$. For $(n, x), (n, x') \in \mathcal{V}_i$, let $B_i^{(n)}[x, x']$ the block matrix component of size $(x_{i-1}(n, x) + 1) \times (x_{i-1}(n, x') + 1)$ whose (j, j')-entry corresponds to the transition rate from $(n, x, j) \in \mathcal{S}_i(n)$ to $(n, x', j') \in \mathcal{S}_i(n)$. The transition of $\mathbf{\Psi}_i$ without changing the level occurs by a service completion at W_{i-2} and blocking of a customer just completed its service at W_i.

(i) *Service completion at* W_{i-2}. Note that if a service is completed at the node W_{i-2}, then the transition of $[\mathbf{\Psi}_i$ from (n, x, j) to $(n, x, j+1)$ occurs, and given $W_{i-1}(t) = (j, k)$, the service at W_{i-2} does not depend on the state (n, x) of $V_i(t)$. Thus the transition rate from (n, x, j) to $(n, x, j+1)$ is $\lambda_{i-1}(j, k)$ and the upper diagonal entries of matrix $B_i^{(n)}[x, x]$ is $\lambda_{i-1}(j, k)$ and others are all zero, that is, the (j, j')-component of $B_i^{(n)}[x, x]$ is for $0 \leq j, j' \leq h$,

$$\left[B_i^{(n)}[x, x]\right]_{jj'} = \lambda_{i-1}(j, y_{i-1}^*(n, x))\mathbf{1}(j' = j+1).$$

(ii) *Occurrence of blocking at* W_i. If a customer at W_i is blocked upon a service completion on the state $\Psi_i(t) = (n, x, j)$, the resulting state of $\Psi_i(t)$ immediately after an occurrence of blocking is $(n, x-1, j)$. Thus the (j, j')-component of $B_i^{(n)}[x, x-1]$ is for $0 \leq j \leq h$ and $0 \leq j' \leq x_{i-1}(n, x-1)$,

$$\left[B_i^{(n)}[x, x-1]\right]_{jj'} = \beta_i(x, n-x)\mathbf{1}(j' = j).$$

The matrices $C_i^{(n)}$. Each component of $C_i^{(n)}$ corresponds to the transition rate from a state in $\mathcal{S}_i(n)$ to a state in $\mathcal{S}_i(n-1)$ which is occurred by a departure of an active customer or a blocked customer from W_i. A departure of an active customer results in the state transition from (n, n, j) to $(n-1, n-1, j)$ for $n \geq 1$ and a departure of a blocked customers results in the state transition from (n, x, j) to $(n-1, x, j)$ for $0 \leq x < n$.

For each $(n,x),(n-1,x')\in\mathcal{V}_i$, denote by $C_i^{(n)}[x,x']$ the block matrix component of size $(h+1)\times(x_{i-1}(n-1,x')+1)$ whose (j,j')-entry corresponds to the transition rate from $(n,x,j)\in\mathcal{S}_i(n)$ and $(n-1,x',j')\in\mathcal{S}_i(n-1)$. It can be easily seen that $C_i^{(n)}[x,x']=0$ except for $x'=x$ with $x<n$ and $x'=n-1$ with $x=n$.

For given $V_i(t)=(n,x)$, let

$$h'=\begin{cases} x_{i-1}(n-1,x), & x<n,\\ x_{i-1}(n-1,n-1), & x=n.\end{cases}$$

It can be easily seen that $h\le h'$. Let $C_{i-1}^u(n,x)$ be the $(h+1)\times(h'+1)$ matrix whose (j,j')-component is the transition probability of $X_{i-1}(t)$ from j to j' given that a departure from W_i is occurred on $\Psi_i(t)=(n,x,j)$. Then it can be easily seen that

$$\left[C_{i-1}^u(n,x)\right]_{jj'}=\mathbf{1}(j'=j),\ 0\le j\le h, 0\le j'\le h'$$

and for $l_i(n)\le x\le u_i(n)$, $l_i(n-1)\le x'\le u_i(n-1)$,

$$C^{(n)}[x,x']=\begin{cases} C_{i-1}^u(n,x)\delta_i(x,n-x)\mathbf{1}(x'=x), & x<n,\\ C_{i-1}^u(n,n)\delta_i(n,0)\mathbf{1}(x'=n-1), & x=n,\\ 0, & \text{otherwise.}\end{cases}$$

4.2. **The subsystem L_i with $b_i=0$.** *The subsystem L_i, $2\le i\le N-1$.* For describing the subsystem L_i with $b_i=0$, define the stochastic processes $\Psi_i^0(t)=(X_i(t),M_i(t),X_{i-1}(t))$. The state space $\mathcal{S}_i^0$ of $\mathbf{\Psi}_i^0=\{\Psi_i^0(t),t\ge0\}$ is $\mathcal{S}_i^0=\cup_{n=0}^{\kappa_i}\mathcal{S}_i^0(n)$, $2\le i\le N-1$, where

$$\mathcal{S}_i^0(n)=\{(n,0,j),(n,0^*,j)\,:\,0\le j\le x_{i-1}(n,n)\}$$

and the number of states in $\mathcal{S}_i^0(n)$ is

$$s_i^0(n)=2(x_{i-1}(n,n)+1),\ 2\le i\le N-1$$

and $y_{i-1}^*(n,n)$ becomes

$$y_{i-1}^*(n,n)=\begin{cases}\max(0,n-c_i), & b_{i-1}>0,\\ 0, & n<c_i, b_{i-1}=0,\\ 0^*, & n=c_i, b_{i-1}=0.\end{cases}$$

The generator of $\mathbf{\Psi}_i^0$ is the same form as (4.1) with

$$\begin{aligned} A_i^{(n)} &= \mathbf{I}_2\otimes A_{i-1}^u(n),\ 0\le n\le\kappa_i-1,\\ & \qquad\qquad (n,0) \qquad (n,0^*)\\ B_i^{(n)} &= \begin{matrix}(n,0)\\(n,0^*)\end{matrix}\begin{pmatrix} B_i^{(n)}[0] & 0\\ \alpha_i(n)\mathbf{I}_{h+1} & B_i^{(n)}[0]\end{pmatrix},\ 0\le n\le\kappa_i,\\ & \qquad\qquad (n-1,0) \qquad (n-1,0^*)\\ C_i^{(n)} &= \begin{matrix}(n,0)\\(n,0^*)\end{matrix}\begin{pmatrix} C_{i-1}^u(n,n)\delta_i^0(n) & C_{i-1}^u(n,n)\beta_i^0(n)\\ 0 & 0\end{pmatrix},\ 1\le n\le\kappa_i,\end{aligned}$$

where $\mathbf{I}_n$ is the identity matrix of size n, $h = x_{i-1}(n,n)$, $h' = x_{i-1}(n+1,n+1)$, and $A^u_{i-1}(n)$ is the $(h+1)\times(h'+1)$ matrix whose (j,j')-component is, for $0 \le j \le h$, $0 \le j' \le h'$,

$$\left[A^u_{i-1}(n)\right]_{jj'} = \mu_{i-1}(j, y^*_{i-1}(n,n))\mathbf{1}(j' = j-1)$$

and $B^{(n)}_i[0]$ is the square matrices of size $h+1$ whose (j,j')-component is, for $0 \le j, j' \le h$,

$$\left[B^{(n)}_i[0]\right]_{jj'} = \lambda_{i-1}(j, y^*_{i-1}(n,n))\mathbf{1}(j' = j+1).$$

The subsystem L_N. Since M_N is never blocked, $Y_N(t) = 0$ and hence $Z_N(t) = X_N(t)$, $V_N(t) = (X_N(t), X_N(t))$, $W_N(t) = (X_N(t), 0)$ and $\kappa_N = \xi_N$. In the later, denote the $Z_N(t)$, $V_N(t)$ and $W_N(t)$ by $X_N(t)$. Thus $\Psi_N(t) = (X_N(t), X_{N-1}(t))$ and the state space of $\mathbf{\Psi}_N$ is

$$\mathcal{S}_N(n) = \{(n,j) \,:\, 0 \le j \le x_{N-1}(n,n)\}.$$

The number of states in $\mathcal{S}_N(n)$ is $s_N(n) = x_{N-1}(n,n)+1$, $0 \le n \le \kappa_N$. The (j,j') component of the matrices $A^{(n)}_N$, $B^{(n)}_N$ and $C^{(n)}_N$ are as follows:

$$\begin{aligned}
\left[A^{(n)}_N\right]_{jj'} &= \mu_{N-1}(j, y^*_{N-1}(n,n))\mathbf{1}(j' = j-1),\\
\left[B^{(n)}_N\right]_{jj'} &= \lambda_{N-1}(j, y^*_{N-1}(n,n))\mathbf{1}(j' = j+1),\\
\left[C^{(n)}_N\right]_{jj'} &= \mu_N(n,0)\mathbf{1}(j' = j).
\end{aligned}$$

5. Approximation of the parameters and performance measures

Now we assume that the system is in stationary state and let

$$\pi_i(n,x,j) = \lim_{t\to\infty} P(\Psi_i(t) = (n,x,j)),$$

and $\boldsymbol{\pi}_i = (\boldsymbol{\pi}_i(n), n = 0,1,\cdots,\kappa_i)$ with $\boldsymbol{\pi}_i(n) = (\pi_i(n,x,j), (n,x,j) \in \mathcal{S}_i(n))$. The limiting distribution $\pi_i(n,s,j)$, $s = 0, 0^*$ of $\Psi^0_i(t)$ and $\boldsymbol{\pi}_i$ are defined similarly.

Performance measures. Once the stationary distribution $\boldsymbol{\pi}_i$ of $\mathbf{\Psi}_i$ is obtained, the performance measures can be obtained as follows:

- Throughput :

$$\Theta = \left(\sum_{n=1}^{\kappa_N} \boldsymbol{\pi}_N(n)\mathbf{e}\right)\mu_N(n,0)$$

- Mean number of customers in W_i :

$$\mathbb{E}[W_i] = \sum_{n=1}^{\kappa_i}\sum_{x=l_i(n)}^{u_i(n)}\sum_{j=0}^{x_{i-1}(n,x)} \min(n, c_i)\boldsymbol{\pi}_i(n,x,j)$$

- Mean number of active customers in W_i :

$$\mathbb{E}[W_i^a] = \sum_{n=1}^{\kappa_i} \sum_{x=l_i(n)}^{u_i(n)} \sum_{j=0}^{x_{i-1}(n,x)} \min(x, a_i)\boldsymbol{\pi}_i(n, x, j)$$

- Mean number of blocked customers in W_i :

$$\mathbb{E}[W_i^b] = \sum_{n=1}^{\kappa_i} \sum_{x=l_i(n)}^{u_i(n)} \sum_{j=0}^{x_{i-1}(n,x)} \min(n - x, b_i)\boldsymbol{\pi}_i(n, x, j).$$

Approximation of $\lambda_i(x, y)$. The marginal distribution $p_i^d(x, y) = P(W_i(t) = (x, y))$ is

$$p_i^d(x, y) = \sum_{j=0}^{x_{i-1}(x+y,x)} \pi_i(x + y, y, j),\ (x, y) \in \mathcal{W}_i,$$

where $\pi_i(x + 0^*, 0^*, j) = \pi_i(x, 0^*, j)$ for $y = 0^*$ in the node with $b_i = 0$. Then $\lambda_i(x, y)$ is approximated by the formula, for $(x, y) \in \mathcal{W}_i$

$$\begin{aligned} \lambda_i(x, y) &= \sum_{j=1}^{h} P(X_{i-1}(t) = j | W_i(t) = (x, y))\mu_{i-1}(j, k) \\ &= \frac{1}{p_i^d(x, y)} \sum_{j=1}^{h} \pi_i(x, y, j)\mu_{i-1}(j, k), \end{aligned}$$

where $h = x_{i-1}(x + y, x)$ and $k = y_{i-1}^*(x + y, x)$.

Approximation of $\beta_{i-1}(x, y)$, $\delta_{i-1}(x, y)$ *for* $b_{i-1} > 0$. We consider two cases of $b_i > 0$ and $b_i = 0$ separately.

Case (i) $b_i > 0$. The marginal distribution $p_{i-1}^u(j, k) = P(W_{i-1}^u(t) = (j, k))$ is given by, $0 \le j \le x_{i-1}^*(k)$

$$\begin{aligned} p_{i-1}^u(j, 0) &= \sum_{(n,x) \in \mathcal{B}_{i-1}(0) \cup \mathcal{D}_{i-1}} \pi_i(n, x, j) = \sum_{n=0}^{c_i} \sum_{x=l_i(n)}^{\min(n,a_i)} \pi_i(n, x, j), \\ p_{i-1}^u(j, k) &= \sum_{(n,x) \in \mathcal{B}_{i-1}(k)} \pi_i(n, x, j) \\ &= \sum_{x=l_i(c_i+k)}^{a_i+k} \pi_i(c_i + k, x, j) + \sum_{n=a_i+k}^{c_i+k-1} \pi_i(n, a_i + k, j),\ 1 \le k \le b_{i-1}. \end{aligned}$$

Thus for $1 \le j \le \xi_{i-1}$,

$$\begin{aligned}
\beta_{i-1}(j,0) &= P(\{V_i(t) \in \mathcal{B}_{i-1}(0) | W_{i-1}(t) = (j,0))\mu_{i-1}(j,0) \\
&= \frac{1}{p^u_{i-1}(j,0)} \left(\sum_{x=l_i(c_i)}^{a_i} \pi_i(c_i, x, j) + \sum_{n=a_i}^{c_i-1} \pi_i(n, a_i, j) \right) \mu_{i-1}(j,0), \\
\delta_{i-1}(j,0) &= \mu_{i-1}(j,0) - \beta_{i-1}(j,0)
\end{aligned}$$

and for $0 \le j \le x^*_{i-1}(k)$, $1 \le k \le b_{i-1}$,

$$\begin{aligned}
\beta_{i-1}(j,k) &= \mu_{i-1}(j,k),\ 0 \le j \le x^*_{i-1}(k), 1 \le k \le b_{i-1}, \\
\delta_{i-1}(j,k) &= \sum_{(n,x) \in \tilde{\mathcal{B}}_{i-1}(k)} P(V_i(t) = (n,x) | W^u_{i-1}(t) = (j,k)) \delta_i(x, n-x) \\
&= \frac{1}{p^u_{i-1}(j,k)} \sum_{(n,x) \in \tilde{\mathcal{B}}_{i-1}(k)} \pi_i(n, x, j) \delta_i(x, n-x)
\end{aligned}$$

with $\delta_N(x,0) = \mu_N(x,0)$ and $\delta_N(x,y) = 0, y \ge 1$.

Case (ii) $b_i = 0$. In case of $b_i = 0$ and $b_{i-1} > 0$, the marginal distribution $p^u_{i-1}(j,k) = P(W^u_{i-1}(t) = (j,k))$ is given by

$$\begin{aligned}
p^u_{i-1}(j,0) &= \sum_{n=0}^{c_i} (\pi_i(n, 0, j) + \pi_i(n, 0^*, j)), \\
p^u_{i-1}(j,k) &= \pi_i(c_i + k, 0, j) + \pi_i(c_i + k, 0^*, j), 1 \le k \le b_{i-1}.
\end{aligned}$$

The $\beta_{i-1}(j,k)$ and $\delta_{i-1}(j,k)$ are approximated as follows:

$$\begin{aligned}
\beta_{i-1}(j,k) &= \mu_{i-1}(j,k),\ 1 \le j \le x^*_{i-1}(k),\ 1 \le k \le b_{i-1} - 1, \\
\beta_{i-1}(j,0) &= P(X_i(t) = c_i | W^u_{i-1}(t) = (j,0)) \mu_{i-1}(j,0) \\
&= \frac{\pi_i(c_i, 0, j) + \pi_i(c_i, 0^*, j)}{p^u_{i-1}(j,0)} \mu_{i-1}(j,0),\ 1 \le j \le \xi_{i-1}, \\
\delta_{i-1}(j,0) &= \mu_{i-1}(j,0) - \beta_{i-1}(j,0),\ 1 \le j \le \xi_{i-1}, \\
\delta_{i-1}(j,k) &= P(W_i(t) = (c_i + k, 0) | W^u_{i-1}(t) = (j,k)) \mu_i(c_i + k, 0) \\
&= \frac{\pi_i(c_i + k, 0, j)}{p^u_{i-1}(j,k)} \mu_i(c_i + k, 0),\ 0 \le j \le x^*_{i-1}(k),\ 1 \le k \le b_{i-1}.
\end{aligned}$$

Approximation of $\alpha_{i-1}(j)$, $\beta^0_{i-1}(j)$ *and* $\delta^0_{i-1}(j)$ *for* $b_{i-1} = 0$.

Case (i) $b_i > 0$. In case of $b_{i-1} = 0$ and $b_i > 0$, the marginal distribution $p^u_{i-1}(j,y) = P(W^u_{i-1}(t) = (j,y))$ is given by

$$\begin{aligned} p^u_{i-1}(j,0^*) &= \sum_{x=l_i(c_i)}^{a_i} \pi_i(c_i,x,j) + \sum_{n=a_i}^{c_i-1} \pi_i(n,a_i,j), \\ p^u_{i-1}(j,0) &= \sum_{n=0}^{c_i-1} \sum_{x=l_i(n)}^{\min(n,a_i-1)} \pi_i(n,x,j). \end{aligned}$$

The formulae for approximation of $\alpha_{i-1}(j)$, $0 \le j \le \xi_{i-1}$ and $\beta^0_{i-1}(j)$, $\delta^0_{i-1}(j)$, $1 \le j \le \xi_{i-1}$ are given as follows:

$$\begin{aligned} \alpha_{i-1}(j) &= P(V_i(t) = (a_i,a_i) | W^u_{i-1}(t) = (j,0^*))\delta_i(a_i,0) \\ &+ \sum_{x=l_i(c_i)}^{a_i-1} P(V_i(t) = (c_i,x) | W_i(t) = (j,0^*))\delta_i(x,c_i-x) \\ &= \frac{1}{p^u_{i-1}(j,0^*)} \left(\pi_i(a_i,a_i,j)\delta_i(a_i,0) + \sum_{x=l_i(c_i)}^{a_i-1} \pi_i(c_i,x,j)\delta_i(x,c_i-x) \right), \\ \beta^0_{i-1}(j) &= \frac{1}{p^u_{i-1}(j,0)} \left(\sum_{x=l_i(c_i-1)}^{a_i-1} \pi_i(c_i-1,x,j) + \sum_{n=a_i-1}^{c_i-2} \pi_i(n,a_i-1,j) \right) \mu_{i-1}(j,0), \\ \delta^0_{i-1}(j) &= \mu_{i-1}(j,0) - \beta^0_{i-1}(j). \end{aligned}$$

Case (ii) $b_i = 0$. In case of $b_i = 0$ and $b_{i-1} = 0$, the marginal distribution of $W^u_{i-1}(t)$ is

$$\begin{aligned} p^u_{i-1}(j,0^*) &= \pi_i(c_i,0,j) + \pi_i(c_i,0^*,j), \\ p^u_{i-1}(j,0) &= \sum_{n=0}^{c_i-1} (\pi_i(n,0,j) + \pi_i(n,0^*,j)). \end{aligned}$$

The formulae for approximation of $\alpha_{i-1}(j)$, $\beta^0_{i-1}(j)$ and $\delta^0_{i-1}(j)$ are given as follows:

$$\begin{aligned} \alpha_{i-1}(j) &= \frac{\pi_i(c_i,0,j)}{p^u_{i-1}(j,0^*)} \mu_i(c_i,0), \ 0 \le j \le \xi_{i-1}, \\ \beta^0_{i-1}(j) &= \frac{\pi_i(c_i-1,0,j) + \pi_i(c_i-1,0^*,j)}{p^u_{i-1}(j,0)} \mu_{i-1}(j,0), \\ \delta^0_{i-1}(j) &= \mu_{i-1}(j,0) - \beta^0_{i-1}(j), \ 1 \le j \le \xi_{i-1}. \end{aligned}$$

6. ALGORITHM

The parameters for the components of Q_i are calculated by the following iterative algorithm.

0. Initial setting.:

(1) Initial assumption: Initially assuming that there are sufficient number of active customers in W_{i-2}, and the customers arrive to the upstream station W_{i-1}^u in the subsystem L_i according to Poisson process with rate $\lambda_{i-1}(j,k) = \mu_{i-2}(M,y)$, where M is a sufficiently large number, for example, $M \geq \max_{1\leq i\leq N} \xi_i$ and $y = y_{i-2}^*(j+k,j)$. For example, if W_{i-2} has m_{i-2} identical servers in parallel whose service time is exponential with rate μ_{i-2}, then

$$\lambda_{i-1}(j,k) = (m_{i-2} - m_{i-2}^b(y))\mu_{i-2},\ 3 \leq i \leq N,$$

where $y = y_{i-2}(j+k,j))$ and

$$m_i^b(y) = \begin{cases} \max(y - b_i^*, 0), & y < b_i \text{ or } y = 0 \text{ with } b_i = 0, \\ m_i, & y = b_i > 0 \text{ or } y = 0^*. \end{cases}$$

Note that $\lambda_1(j,k)$ is determined by the assumption of arrival process.

(2) Construct the matrices $A_i^{(n)}$, $2 \leq i \leq N$. Note that the matrices $A_i^{(n)}$ do not contain any unknown parameters and they are not necessary to be updated in iteration step.

(3) Calculate $\boldsymbol{\pi}_N$ of Q_N and throughput $\Theta^{(0)}$, and calculate $\beta_{N-1}(x,y)$ and $\delta_{N-1}(x,y)$ for $b_{N-1} > 0$ and $\alpha_{N-1}(x)$, $\beta_{N-1}^0(x)$ and $\delta_{N-1}^0(x)$ for $b_{N-1} = 0$. Note that the matrices $C_N^{(n)}$ are not necessary to be updated in the iteration step.

1. Backward step.: For $i = N-1, N-2, \cdots, 2$,

(1) update the matrices $B_i^{(n)}$ and $C_i^{(n)}$ using $\beta_i(x,y)$ and $\delta_i(x,y)$ for $b_{i-1} > 0$ ($\alpha_i(x)$, $\beta_i^0(x)$ and $\delta_i^0(x)$ for $b_{i-1} = 0$) calculated in the previous step and calculate $\boldsymbol{\pi}_i$ of Q_i, and

(2) if $i \geq 3$, then update $\beta_{i-1}(x,y)$ and $\delta_{i-1}(x,y)$ for $b_{i-1} > 0$ and $\alpha_{i-1}(x)$, $\beta_{i-1}^0(x)$ and $\delta_{i-1}^0(x)$ for $b_{i-1} = 0$ using the formulae in Section 5.

(3) If $i = 2$, then update $\lambda_2(x,y)$ using the stationary distribution $\boldsymbol{\pi}_2$ of Q_2, and go to the forward step.

2. Forward step.: For $i = 3, \cdots, N-1$,

(1) update the matrices $B_i^{(n)}$ using $\lambda_{i-1}(x,y)$ calculated in the previous step and calculate $\boldsymbol{\pi}_i$ of Q_i, and

(2) update $\lambda_i(x,y)$ using the formulae in Section 5.

3. Tolerance check.: In the last subsystem L_N, calculate the throughput and check the stopping criterion as follows:

(1) Update $B_N^{(n)}$ and $\boldsymbol{\pi}_N$ of Q_N.

(2) Calculate throughput $\Theta^{(m)}$.

(3) Check the tolerance (stopping criterion)

$$TOL = |\Theta^{(m)} - \Theta^{(m-1)}| < \epsilon, \tag{6.1}$$

where $\Theta^{(m)}$ is the throughput obtained in the mth iteration and $\epsilon > 0$ is the predetermined tolerance. If the stopping criterion is satisfied, then stop the iteration, otherwise

update $\beta_{N-1}(x,y)$ and $\delta_{N-1}(x,y)$ for $b_{N-1}>0$ ($\alpha_{N-1}(x)$, $\beta^0_{N-1}(x)$ and $\delta^0_{N-1}(x)$ for $b_{N-1}=0$), and repeat the backward and forward iterations until the stopping criterion is satisfied.

Remark. One can start the iteration with initial guess of the departure rates under the assumption that W_i^d in L_i is never blocked. In this case, the iteration step is performed by the procedure that the arrival rates are updated in forward step, and then departure rates are updated in backward step in the first iteration, and repeat this procedure until the stopping criterion is satisfied.

Complexity of algorithm. The stationary distribution $\boldsymbol{\pi}_i$ of Q_i can be calculated using well-known matrix geometric method (e.g. Shin [14]). Within the iterative algorithm, solving a subsystem is time consuming. To solve subsystem L_i using the algorithm, we must invert $K_i = \kappa_i + 1$ matrices with the maximum size $s_i^* = \max_{0\le n\le K_i} s_i(n)$. Therefore the complexity of one iteration becomes $O(\sum_{i=1}^N K_i (s_i^*)^3)$.

The number of iterations required is difficult to predict because it depends on the tolerance ϵ and the length of the line and system parameters. For example, as shown from numerical experiments in Section 8, the number of iterations increases with the line length. Although the convergence of the iteration scheme is not proven analytically, extensive numerical experiments indicate the convergence of the iteration.

7. TANDEM QUEUES WITH MULTIPLE SERVERS UNDER GENERAL BLOCKING SCHEME

We apply the method to the system with multiple exponential servers and several blocking scheme.

Consider a tandem queue in which the node W_i consists of service station M_i and a buffer G_i of capacity g_i as depicted in Fig.1. The service station M_i has m_i identical servers in parallel and the service time of each server at M_i is of exponential with rate μ_i. The capacity of the node W_i is $c_i = g_i + m_i$. Let a_i^* be the size of buffer space in G_i for active customers and let $b_i^*(\le b_i)$ be the size of buffer space in G_i for blocked customers. The node W_i can contain $a_i = a_i^* + m_i$ active customers. Note that $0 \le a_i^* \le g_i$, $0 \le b_i^* \le g_i$ and $a_i^* + b_i^* \ge g_i$ and the maximal number of blocked servers is $b_i^{**} = b_i - b_i^*(\le m_i)$. Assuming that the last node M_N is never blocked, $b_N = 0$ and $a_N = c_N$.

The source node W_0 behaves like a system with m_0 servers and a virtual buffer of size b_0 for blocked customers to enter the first node W_1. We assume that W_0 is never starved and each server in W_0 starts service immediately after a service completion unless the server is blocked. The service time of each server is exponential with rate μ_0. The arrival rate to W_1 is

$$\lambda_1(x,y) = (m_0 - y_0^b(x,y))\mu_0,$$

where $y_0^b(x,y)$ is the number of blocked servers in W_0. If W_0 is a BBS node, then the customers arrive according to an ordinary Poisson process with constant rate $m_0\mu_0$ and blocked customers are lost, that is,

$$\lambda_1(x,y) = \begin{cases} m_0\mu_0, & x < a_1, x + y < c_1, \\ 0, & \text{otherwise.} \end{cases}$$

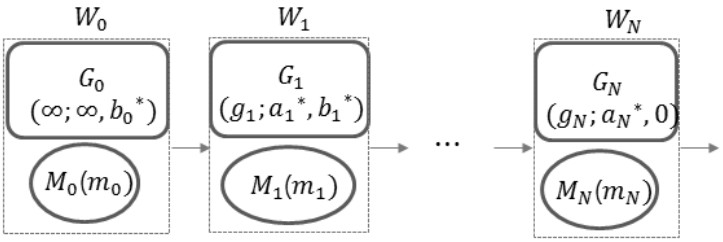

FIGURE 1. Tandem queueing network with general blocking

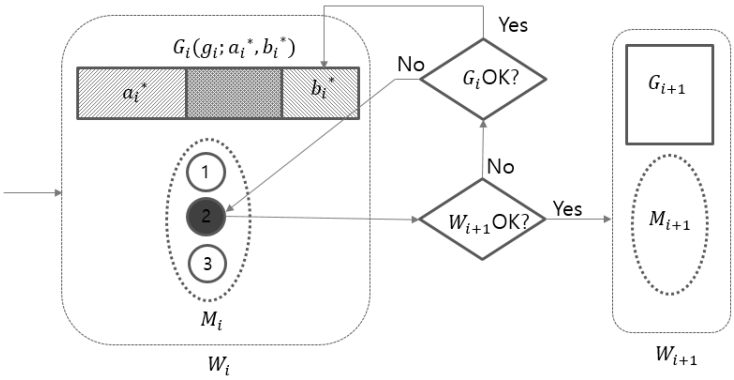

FIGURE 2. Blocking mechanism for $b_i > 0$

Case (i) *BAS node* $(b_i > 0)$. In this case, the blocked customer is stocked at the buffer G_i if there ia a place available for blocked customers, otherwise it is stayed at the server just service completed and the server is blocked (Figure 2).

The server blocked at M_i is forced to stop its service until there is a place available for active customers at W_{i+1}. If the number of blocked customers at W_i reaches b_i upon a service completion, then all the servers at M_i are blocked and forced to stop their service. We coin this type of blocking rule the *generalized blocking after service (GBAS)* rule. This blocking scheme contains many BAS blocking schemes as a special cases:

(1) The blocking scheme with $b_i = b_i^* + m_i$ is an ordinary GB scheme in the system with multiple servers.
(2) If $b_i^* = 0$ and $b_i = m_i$, then the blocking mechanism is the ordinary BAS rule in [12].
(3) If $b_i = b_i^* > 0$, then the blocked customers are stocked only at the buffer. All the servers are blocked upon the blocking level reaches b_i, however, the idle server can accept an active customer even it is stopped its service. The blocking state of all the servers changes to working states if the level of blocked customers downs to the below of b_i.

(4) When $0 < b_i^{**} < m_i$, then all the servers at M_i are forced to stop their service upon the number of the blocked customers at M_i reaches b_i, however the idle server can accept an active customer even it is stopped its service. The servers that are not blocked start new service whose length is exponential distributed random variable with rate μ_i upon the level of blocked customers becomes less than b_i.

Given $W_i(t) = (x, y)$, the number of blocked servers $m_i^b(y)$ at M_i is

$$m_i^b(y) = \begin{cases} \max(y - b_i^*, 0), & y < b_i, \\ m_i, & y = b_i \end{cases}$$

and hence the service rate from M_i is given by

$$\mu_i(x, y) = \min(x, m_i - m_i^b(y))\mu_i.$$

Case (ii) *BBS node* $(b_i = 0)$. In this case, the service rate of M_i depends on the number x of AC and the server state s of $M_i(t)$ and is given by

$$\mu_i(x, s) = \begin{cases} \min(x, m_i)\mu_i, & s = 0, \\ 0, & s = 0^*. \end{cases}$$

8. Numerical Results

To investigate the accuracy of the method proposed in this paper, the method is applied to the tandem queue with multiple servers and the results are compared with simulations. The simulation models for the systems in the tables were developed using ARENA [15]. The simulation run time was set to 100,000 unit times, including a warm-up time of 10,000 unit times. Ten replications were conducted for each case and the half length of the 95% confidence interval (c.i.) was obtained. A tolerance $\epsilon = 10^{-5}$ is used for stopping criterion in (6.1). To validate the simulation program, the simulation are compared with the exact one for throughput of the system with $N = 2$ in Table 1. The table shows that simulation can be used as an alternative of exact analysis.

The current approximation for the ordinary BAS system with $N = 5$, $N = 11$ and $b_i^* = 0$, $b_i = m_i$ is compared with the method (SM) of Shin and Moon [12] in Tables 2. The simulation

TABLE 1. Comparison of simulation with exact results for throughput of the system with $N = 2$.

m_i	g_i	(b_0^*, b_1^*, b_2^*)	(b_0, b_1, b_2)	Exact	Sim (c.i.)
1	3	(0,0,0)	(0,0,0)	0.7359	0.7357 ($\pm$0.0017)
		(0,0,0)	(0,1,1)	0.7543	0.7542 ($\pm$0.0012)
		(0,2,2)	(0,3,3)	0.7677	0.7675 ($\pm$0.0014)
		(3,2,2)	(3,3,3)	0.8266	0.8257 ($\pm$0.0012)
3	3	(0,0,0)	(0,0,0)	0.7744	0.7798 ($\pm$0.0011)
		(0,0,0)	(0,3,3)	0.7966	0.8009 ($\pm$0.0008)
		(0,0,0)	(3,3,3)	0.8360	0.8363 ($\pm$0.0009)
		(3,3,3)	(3,5,5)	0.8495	0.8520 ($\pm$0.0009)

results (Sim) and the numerical results for SM in Tables 3 are from [12]. The measure of the deviation of approximation (App) from the simulation (Sim) is calculated by the formula $\mathcal{D}(\%) = ((\text{App} - \text{Sim})/\text{Sim}) \times 100$. Table 3 shows that the accuracy of current method is similar to that of SM.

We consider a tandem queue that consists of N nodes and each node has multiple servers and follows a general blocking scheme. Customers arrive from outside (W_0) according to a Poisson process and the blocked customers entering into the node W_1 are lost, that is, $m_0 = 1$, $b_0 = 0$, and $\mu_0 = 1.0$. We assume that $a_i = m_i + g_i$ and $a_i^* = g_i$. In the following, the N-dimensional vectors $\boldsymbol{m} = (m_1, \cdots, m_N)$ and $\boldsymbol{g} = (g_1, \cdots, g_N)$ mean the number of servers and the buffer size in the system, respectively. For example, the vector $\boldsymbol{m} = (1, 3, 2, 4)$ means that $m_1 = 1$, $m_2 = 3$, $m_3 = 2$, $m_4 = 4$. Similarly, denote the upper limits of customers in buffer and in the system by the N-dimensional vectors $\boldsymbol{b}^* = (b_1^*, \cdots, b_{N-1}^*, 0)$ and $\boldsymbol{b} = (b_1, \cdots, b_{N-1}, 0)$, respectively. If the number m_i of servers at node W_i, $i = 1, 2, \cdots, N$ are the same as a constant k, we write $m_i = k$ instead of using vector $\boldsymbol{m}$. Similarly, if g_i, $1 \le i \le N$ are the same as a constant j, then we write j in stead of vector $\boldsymbol{g}$.

TABLE 2. Throughput for the system of length $N = 5$ with $g_i = 3$

	$\boldsymbol{m}$	$(3,3,3,3,3)$		$(3,1,3,1,3)$	
	μ_i	$\frac{1}{m_i}$		1.0	
b_i^*	b_i	Sim	App ($\mathcal{D}(\%)$)	Sim	App ($\mathcal{D}(\%)$)
0	0	0.727	0.716 (-1.6)	0.862	0.857 (-0.6)
m_i	m_i	0.764	0.762 (-0.2)	0.869	0.870 (0.1)
3	3	0.764	0.771 (0.9)	0.877	0.877 (0.0)
2	3	0.777	0.774 (-0.4)	0.877	0.877 (-0.1)
m_i	c_i	0.786	0.782 (-0.4)	0.879	0.878 (-0.1)

TABLE 3. Throughput for tandem queues under ordinary BAS blocking

N	m_i	μ_i	g_i	Sim(CI)	SM ($\mathcal{D}(\%)$)	App ($\mathcal{D}(\%)$)
5	3	$\frac{1}{m_i}$	0	0.643 ($\pm$0.002)	0.634 (-1.3)	0.637 (-0.8)
			3	0.776 ($\pm$0.003)	0.777 (0.1)	0.775 (0.0)
			5	0.819 ($\pm$0.004)	0.820 (0.1)	0.819 (0.0)
		1.0	0	1.930 ($\pm$0.003)	1.903 (-1.4)	1.912 (-0.9)
			3	2.330 ($\pm$0.004)	2.330 (0.0)	2.326 (-0.2)
			5	2.461 ($\pm$0.005)	2.461 (0.0)	2.458 (-0.1)
11	$m_{2i} = 4$,	$\frac{1}{m_i}$	0	0.557 ($\pm$0.002)	0.545 (-1.7)	0.543 (-2.6)
	$m_{2i+1} = 1$,		3	0.734 ($\pm$0.001)	0.737 (0.4)	0.730 (-0.4)
	$i = 0, 1, \cdots, 5$		5	0.788 ($\pm$0.003)	0.790 (0.2)	0.786 (-0.3)
		1.0	0	0.726 ($\pm$0.003)	0.725 (-0.2)	0.721 (-0.7)
			3	0.853 ($\pm$0.002)	0.863 (1.2)	0.854 (-0.1)
			5	0.888 ($\pm$0.002)	0.894 (0.7)	0.889 (-0.1)

The throughput for the system of length $N = 5$ with $g_i = 3$ are listed in Table 2. The throughput and the mean number $\mathbb{E}[W] = \sum_{i=1}^{N} \mathbb{E}[W_i]$ of customers in the system are presented in Table 5 for the systems of length $N = 10$ with parameters in Table 4. The mean number $\mathbb{E}[W_i]$ of customers in node W_i are presented in Table 6. In Table 2, Table 5 and Table 6, the half length of 95% confidence interval of simulation results for throughput and for the mean number of customers are less than 0.002 and 0.3, respectively, and confidence intervals are omitted. The table shows that the approximation works well for the throughput and mean number of customers.

TABLE 4. Scenarios for the system of length $N = 10$ in Table 5 and Table 6

Cases	Type	m_i	g_i	b_i^*	b_i
1	BBS	1	3	0	0
2	BAS	1	3	2	2
3	BAS	1	3	2	3
4	Kanban	1	3	3	4
5	Mixed	1	3	$\boldsymbol{b}_1^*$	$\boldsymbol{b}_1$
6	BBS	$\boldsymbol{m}$	$\boldsymbol{g}$	0	0
7	BAS	$\boldsymbol{m}$	$\boldsymbol{g}$	$\boldsymbol{g}_0$	$\boldsymbol{g}_0$
8	Kanban	$\boldsymbol{m}$	$\boldsymbol{g}$	$\boldsymbol{g}_0^*$	5
9	Mixed	$\boldsymbol{m}$	$\boldsymbol{g}$	$\boldsymbol{b}_2^*$	$\boldsymbol{b}_2$
10	Mixed	$\boldsymbol{m}$	$\boldsymbol{g}$	$\boldsymbol{b}_3^*$	$\boldsymbol{b}_3$

$\boldsymbol{m} = (4, 1, 3, 2, 3, 4, 1, 3, 1, 2)$
$\boldsymbol{g} = (1, 4, 2, 3, 2, 1, 4, 2, 4, 3)$
$\boldsymbol{g}_0 = (1, 4, 2, 3, 2, 1, 4, 2, 4, 0)$
$\boldsymbol{b}_1^* = (2, 2, 0, 2, 2, 2, 0, 2, 2, 0)$
$\boldsymbol{b}_1 = (3, 3, 0, 3, 3, 3, 0, 3, 3, 0)$
$\boldsymbol{b}_2^* = (1, 3, 2, 3, 0, 0, 4, 2, 3, 0)$
$\boldsymbol{b}_2 = (3, 4, 4, 5, 3, 2, 5, 5, 4, 0)$
$\boldsymbol{b}_3^* = (1, 2, 0, 0, 1, 1, 0, 1, 4, 0)$
$\boldsymbol{b}_3 = (5, 3, 0, 2, 4, 3, 0, 3, 5, 0)$

TABLE 5. Throughput and mean number $\mathbb{E}[W]$ of customers in the system with $N = 10$

		Throughput		$\mathbb{E}[W]$	
Cases	μ_i	Sim	App ($\mathcal{D}(\%)$)	Sim	App ($\mathcal{D}(\%)$)
1	1.0	0.641	0.637 (-0.6)	20.1	20.0 (-0.5)
2	1.0	0.714	0.710 (-0.5)	24.2	24.0 (-0.9)
3	1.0	0.730	0.724 (-0.7)	24.9	24.6 (-1.3)
4	1.0	0.737	0.731 (-0.8)	25.1	24.8 (-1.2)
5	1.0	0.703	0.697 (-0.8)	24.1	23.8 (-1.2)
6	$\frac{1}{m_i}$	0.662	0.645 (-2.56)	28.1	27.0 (-3.9)
	1.0	0.871	0.881 (1.1)	26.6	25.6 (-3.6)
7	$\frac{1}{m_i}$	0.716	0.710 (-0.8)	30.0	29.9 (-0.3)
	1.0	0.890	0.895 (0.6)	26.2	25.6 (-2.4)
8	$\frac{1}{m_i}$	0.730	0.726 (-0.6)	30.8	30.7 (-0.2)
	1.0	0.890	0.890 (0.0)	26.0	26.1 (0.2)
9	$\frac{1}{m_i}$	0.721	0.720 (-0.1)	30.6	30.8 (0.5)
	1.0	0.890	0.892 (0.2)	25.8	25.6 (-0.6)
10	$\frac{1}{m_i}$	0.704	0.699 (-0.7)	30.7	30.3 (-1.4)
	1.0	0.877	0.889 (1.3)	27.7	28.0 (1.1)

Run time. The current algorithm was performed on a laptop computer at 2.80GHz 16.0 GB RAM using Mathematica®11 [16] for the system with $a_i = b_i = c_i$, $a_i^* = b_i^* = g_i$ and $\mu_i = \frac{1}{m_i}$. The stopping criterion $\epsilon = 10^{-5}$ was used. The number of iterations (NI) and run time (CPU) in seconds are listed in Table 7. The behavior of the run time and the number of iteration of the algorithm as a function of the buffer size is depicted in Fig. 3. The table and figure show that the run time of the algorithm increases with the line length, buffer size and the number of servers at each node, and it depends significantly on the buffer size and the number of servers. The number of iterations is more sensitive to the length N of the line than the buffer size and the number of servers.

TABLE 6. Mean number $\mathbb{E}[W_i]$ of customers at node W_i in the system with $N = 10$, $\mu_i = \frac{1}{m_i}$

	Case 1 (BBS)		Case 4 (Kanban)		Case 9 (Mixed)	
Node	Sim	App ($\mathcal{D}(\%)$)	Sim	App ($\mathcal{D}(\%)$)	Sim	App ($\mathcal{D}(\%)$)
1	2.68	2.69 (0.2)	2.36	2.36 (0.3)	3.55	3.55 (0.1)
2	2.42	2.43 (0.2)	2.67	2.67 (-0.2)	3.25	3.22 (-1.0)
3	2.28	2.27 (-0.4)	2.74	2.72 (-0.6)	3.91	3.86 (-1.1)
4	2.16	2.15 (-0.5)	2.74	2.71 (-1.2)	3.04	2.85 (-6.3)
5	2.06	2.05 (-0.5)	2.71	2.66 (-1.9)	3.44	3.34 (-2.7)
6	1.96	1.95 (-0.7)	2.66	2.60 (-2.2)	3.72	3.64 (-2.1)
7	1.86	1.84 (-0.9)	2.58	2.53 (-2.1)	2.87	2.90 (1.3)
8	1.75	1.73 (-1.2)	2.46	2.42 (-1.8)	2.71	2.69 (-0.8)
9	1.59	1.57 (-1.1)	2.27	2.25 (-1.3)	2.03	2.02 (-0.1)
10	1.32	1.31 (-1.1)	1.92	1.90 (-1.3)	2.23	2.22 (-0.5)
Total	20.09	19.99 (-0.5)	25.12	24.81(-1.2)	30.74	30.30 (-1.4)

TABLE 7. CPU time for Kanban system in seconds ($\mu_i = 1.0/m_i$)

		$m_i = 1$		$m_i = 3$		$m_i = 5$	
N	g_i	NI	CPU	NI	CPU	NI	CPU
10	1	8	0.2	7	1.4	7	4.8
	3	7	1.4	7	4.9	7	13.7
	5	7	4.9	7	13.8	7	32.0
	7	7	13.6	7	31.8	7	69.5
20	1	15	1.0	14	6.1	14	22.9
	3	13	5.6	13	20.5	13	60.6
	5	12	19.6	12	56.1	12	133.2
	7	12	55.8	12	131.4	12	284.4

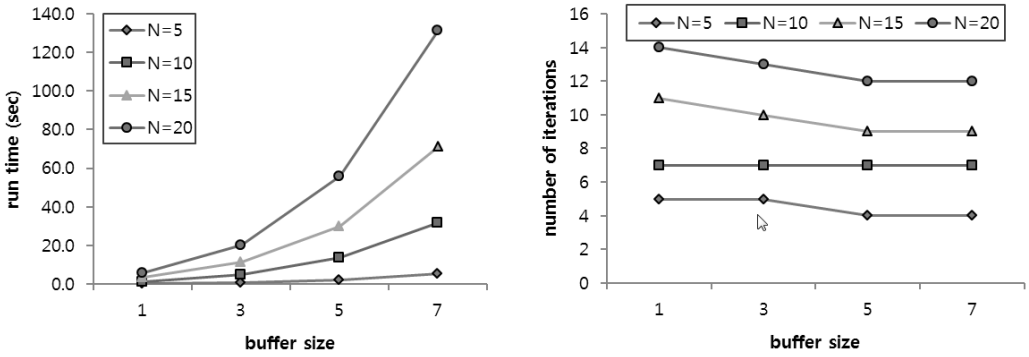

FIGURE 3. Run time and the number of iterations of algorithm for kanban system with $m_i = 3$ and $\mu_i = \frac{1}{m_i}$

9. CONCLUDING REMARKS

An approximation method for tandem queues with finite buffers and exponential service times with state dependent service rates under general blocking scheme was presented. The model considered in this paper is very general and it contains the system with multi-server node and many classical blocking scheme such as ordinary manufacturing blocking, communication blocking, and kanban blocking as special cases. Extensive numerical experiments show that the current method is very effective in the sense of accuracy of approximation and computation time even for the system that consists of nodes with different blocking schemes.

REFERENCES

[1] Y. Dallery and S. B. Gershwin, *Manufacturing flow line systems: a review of models and analytical results*, Queueing Systems, **12** (1992), 3–94.

[2] J. Li, D. E. Blumenfeld, N. Huang and J. M. Alden, *Throughput analysis of production systems: recent advances and future topics*, International Journal of Production Research, **47** (2009), 3823–3851.

[3] H. T. Papadopoulos and C. Heavey, *Queueing theory in manufacturing systems analysis and design: a classification of models for production and transfer lines*, European Journal of Operational Research, **92** (2016), 1–27.

[4] S. Balsamo, V. de Nitto Personé and R. Onvural, *Analysis of Queueing Networks with Blocking*, Kluwer Academic Publishers, Dordrecht, 2001.

[5] D. W. Cheng and D. D. Yao, *Tandem queues with general blocking: a unified model and comparison results*, Discrete Event Dynamic Systems: Theory and Applications, **2** (1993), 207–234.

[6] D. W. Cheng, *Analysis of a tandem queue with state dependent general blocking: a GSMP perspective*, Performance Evaluation, **17** (1993), 169–173.

[7] P. Glasserman and D. D. Yao, *Structured buffer-allocation problems*, Discrete Event Dynamic Systems: Theory and Applications, **6** (1996), 9–41.

[8] S. Balsamo, *Queueing Networks with Blocking: Analysis, solution algorithms and properties*, Next Generation Internet, D. Kouvatsos (Ed.), LNCS 5233, Springer-Verlag, 2011.

[9] J. A. Buzacott and J. G. Shanthikumar, *Design of manufacturing systems using queueing models*, Queueing Systems, **12** (1992), 135–213.
[10] R. Bierbooms, I.J.B.F. Adan and M. van Vuuren, *Approximate analysis of single-server tandem queues with finite buffers*, Annals of Operations Research, **209** (2013), 67–84.
[11] S. B. Gershwin, *Manufacturing Systems Engineering*, American Studies, Prentice-Hall, 1994.
[12] Y. W. Shin and D. H. Moon, *Approximation of throughput in tandem queues with multiple servers and blocking*, Applied Mathematical Modelling, **38** (2014), 6122–6132.
[13] Y. W. Shin and D. H. Moon,*A unified approach for an approximation of tandem queues with failures and blocking under several types of service-failure interactions*, Computers and Operations Research, **127** (2021), 1-16.
[14] Y. W. Shin, *Fundamental matrix of transition QBD generator with finite states and level dependent transitions*, Asia-Pacific Journal of Operational Research, **26** (2009), 1–18.
[15] W. D. Kelton, R. P. Sadowski and D. A. Sadowski, *Simulation with ARENA*, 2^{nd} edition. McGraw-Hill, New York, 1998.
[16] S. Wolfram, *Mathematica*, 2nd ed. Addison-Wesley, 1991.