# Approximation of Tandem Queues with Blocking

Dug Hee Moon[1] and Yang Woo Shin[2,*]
[1]*School of Industrial Engineering and Naval Architecture, Changwon National University, Changwon, Gyeongnam 51140, Korea*
[2]*Department of Statistics, Changwon National University, Changwon, Gyeongnam 51140, Korea*

Keywords: Tandem Queue, Phase Type Distribution, Decomposition, Blocking.

Abstract: In this paper, we present an approximate analysis for tandem queues with single reliable server at each service station and a buffer of finite capacity between service stations. Blocking-After-Service (BAS) rule is adopted. The effects of the moments of service times to the throughput are investigated numerically and the service time is fitted with a phase type (PH) distribution by matching the first two moments. The system with phase type service times is approximated based on the decomposition method with two-server-one-buffer subsystem. Some numerical examples are presented for accuracy of approximation.

## 1 INTRODUCTION

Tandem queue sometimes called transfer line in manufacturing system with finite buffers have been widely used for performance modeling of computer systems and production systems. e.g. see the monographs Gershwin (1994), Buzzacott and Shanthikumar (1993), the survey papers Dallery and Gershwin (1992), Papadopoulos and Heavey (1996) and the references therein. Although the system with finite buffer is modeled by a Markov chain, the number of states of the Markov chain increases drastically as the number of stages increases, which makes analytical or numerical solutions intractable for the systems with long line. Approximations of the queueing networks have been developed in many directions. One is to overcome the problem of dimension of state space and another is to reduce the assumption of exponential service time. The system with phase type service time or approximate formula of $G/G/m/N$ system have been used for approximate analysis of the system with non-exponential service time. One of the most common method among the approximation techniques to solve the dimensional problem is decomposition method developed by Gershwin (1987,1994). The method decomposes the long line into subsystems with two service stations and one buffer, and derives a set of equations that determine the unknown parameters of each subsystem, and finally develops an iterative algorithm to solve these equations. There are some approximations for the system that the service time distributions are not exponential or geometric (in discrete time case) based on decomposition method, see Templemeier and Bürger (2001), Bierbooms et al. (2011) for system with general service times, Helber (2005) for the reliable systems with Cox-2 distribution of service time, and Colledani and Tolio (2011), Shin and Moon (2018) for discrete time system with unreliable servers of discrete PH-distribution of geometric or repair time.

In this paper, we present an approximate analysis for tandem queues with single reliable server at each service station and a buffer of finite capacity between service stations under the blocking-after-service rule. The approximation is based on the decomposition method with two-server-one-buffer subsystem. The system with phase type (PH) distribution of service time is approximated, where the states of the server include the state of the number of customers in upstream subsystem as well as the states of the server (blocking, starvation, working) to reflect the dependence of consecutive subsystems. In case of general distribution of service times, approximate the distribution of service times of original system with PH-distributions and then use the system with PH-service time as an approximation of original system.

The paper is organized as follows. The effects of moments of service times to the throughput are numerically investigated in Section 2. The approximation procedure is described in Section 3. An algorithm for the parameters of subsystems is presented in Section 4. The effectiveness of the approximation is investigated numerically in Section 5. Finally, concluding remarks are given in Section 6.



Moon, D. and Shin, Y.
Approximation of Tandem Queues with Blocking.
DOI: 10.5220/0007469504220428
In *Proceedings of the 8th International Conference on Operations Research and Enterprise Systems (ICORES 2019)*, pages 422-428
ISBN: 978-989-758-352-0

## 2 SENSITIVITY OF THROUGHPUT

In this section, the effects of moments of service times to the throughput are numerically investigated. We consider the systems with five servers where the service times are identical with common means 1.0 and the buffer size between servers are identical. The buffer size $c_i$ between servers are chosen as $c_i = 1, 5, 10$. We use the Erlang distribution of order $k$ $(E_k)$ for squared coefficient of variation $C_s^2 = \frac{1}{k} < 1$, exponential distribution $(Exp)$ for $C_s^2 = 1$ and hyperexponential distribution of order 2 $(H_2)$ with balanced mean for $C_s^2 > 1$ whose probability density function is

$$f(t) = p_1\lambda_1 e^{-\lambda_1 t} + p_2\lambda_2 e^{-\lambda_2 t},\ t \geq 0,$$

with $\lambda_1 = 2p_1\mu$, $\lambda_2 = 2p_2\mu$ and

$$p_1 = \frac{1}{2}\left(1 + \sqrt{\frac{C_s^2 - 1}{C_s^2 + 1}}\right),\ p_2 = 1 - p_1.$$

The lognormal distribution (LN), gamma distribution (GAM) and Weibul distribution (WEIB) are compared with the PH-distributions whose the first two moments are the same as those of the distributions LN, GAM and WEIB. Simulation results of the throughput are listed in Table 1. The deviation $(\mathcal{D})$ between PH-distribution and the non PH-distribution (G) is calculated by

$$\mathcal{D}(\%) = \frac{\text{PH} - \text{G}}{\text{PH}} \times 100,$$

where PH and G are the simulation results for throughput of PH-distribution and others, respectively. The table shows that the deviation decreases as the buffer size increases and the deviation is sufficiently small in practical sense even for the system with moderate buffer size $(c_i = 5)$ and high variability of service time $(C_s^2 = 5.0)$. We have observed the similar phenomena for the system with 10 servers, but the results are not listed here. Based on the observations, we approximate the system with general distribution of service time using the system with PH-service time.

Table 1: Throughput for tandem queues with five servers.

| $C_s^2$ | $c_i$ Serv. | 1 sim($\mathcal{D}$(%)) | 5 sim($\mathcal{D}$(%)) | 10 sim($\mathcal{D}$(%)) |
|---|---|---|---|---|
| 0.25 | $E_4$ | 0.805 | 0.929 | 0.961 |
| | LN | 0.803(-0.2) | 0.928(-0.1) | 0.960(+0.0) |
| | WEIB | 0.808(+0.4) | 0.930(+0.1) | 0.961(+0.0) |
| 0.5 | $E_2$ | 0.712 | 0.875 | 0.927 |
| | LN | 0.713(+0.2) | 0.873(-0.2) | 0.926(-0.1) |
| | WEIB | 0.713(+0.2) | 0.876(+0.1) | 0.927(+0.0) |
| 1.0 | Exp | 0.608 | 0.793 | 0.869 |
| | LN | 0.618(+1.7) | 0.795(+0.2) | 0.869(-0.1) |
| 2.0 | $H_2$ | 0.513 | 0.690 | 0.784 |
| | LN | 0.530(+3.3) | 0.701(+1.6) | 0.789(+0.7) |
| | GAM | 0.508(-1.0) | 0.689(+0.0) | 0.785(+0.1) |
| | WEIB | 0.508(-1.0) | 0.689(-0.1) | 0.784(+0.0) |
| 5.0 | $H_2$ | 0.416 | 0.553 | 0.646 |
| | LN | 0.416(-0.0) | 0.553(+0.0) | 0.646(+0.0) |
| | GAM | 0.408(-1.8) | 0.558(+0.9) | 0.655(+1.4) |
| | WEIB | 0.402(-3.3) | 0.549(-0.7) | 0.645(-0.1) |

## 3 APPROXIMATION OF TANDEM QUEUE WITH *PH*-SERVICE TIME

### 3.1 Model

We consider a tandem queueing network $L$ with $N^* = N + 1$ servers $M_i$, $i = 0, 1, \cdots, N$ and buffers $B_i$ of capacity $c_i$ with $0 \leq c_i < \infty$ between $M_{i-1}$ and $M_i$, $i = 1, 2, \cdots, N$. Service time distribution of $M_i$ is of phase type denoted by $PH(\boldsymbol{\alpha}_i, S_i)$, where $\boldsymbol{\alpha}_i = (\alpha_i(1), \cdots, \alpha_i(h_i))$ is a probability distribution and $S_i = (s_i(i,j))$ is a nonsingular matrix of size $h_i$ such that $s_i(j,k) \geq 0$, $j \neq k$, $s_i(j,j) = -s_i(j) < 0$, $1 \leq j \leq h_i$. Let $\boldsymbol{S}_i^0 = -S_i\mathbf{e} = (s_i^0(1), \cdots, s_i^0(h_i))^t$, where $\mathbf{e}$ is the column vector of appropriate size whose components are all 1. See Neuts (1981) for more about phase type distribution. Transportation times of customers through buffers and servers are assumed to be negligible comparing to service time. We assume the blocking after service (BAS) rule. That is, when a server completes its service at a stage, if the buffer of next stage is full at that time, then the server is forced to stop its service and the customer is held at the station where it just completed its service until the destination can accommodate it. A server $M_i$ is said to be starved if there are no customers to be served on the server $M_i$. We assume that the initial server $M_0$ is never starved and it starts new service immediately after a service completion unless the server is blocked. The last server $M_N$ is never blocked and the customer at $M_N$ leaves the system immediately after completing its service.

Denote the state of the server $M_i$ at time $t$ by

$$M_i(t) = \begin{cases} j, & M_i \text{ is working and its service phase is } j \\ b, & M_i \text{ is blocked} \\ s, & M_i \text{ is starved} \end{cases}$$

and let $B_i(t)$ be the number of customers waiting in $B_i$ at time $t$. By $X_i(t)$, denote the number of customers waiting in $B_i$ and the customers being served or blocked at $M_i$ and the customers blocked at $M_{i-1}$

at time $t$, that is,

$$X_i(t) = B_i(t) + 1(M_{i-1}(t) = b) + 1(M_i(t) \in \{\boldsymbol{w}(i), b\}),$$

where $1(A) = 0$ if $A$ is true and 0, otherwise and $\boldsymbol{w}(i) = \{1, 2, \cdots, h_i\}$ the set of phases of service time of $M_i$. Then $X_i(t)$ takes values on $\{0, 1, \cdots, K_i\}$, where $K_i = c_i + 2$.

In this paper, we propose an approximate analysis of this system based on decomposition approach. The first step is to decompose the $N+1$ server system into a set of $N$ two-server-one-buffer subsystems $L_i$, $i = 1, 2, \cdots, N$. Each subsystem $L_i$ is consists of upstream server $M_{i-1}$, downstream server $M_i$ and a buffer $B_i$ between them. The subsystems are modeled with a quasi-birth-and-death process. The transition rates among the states of servers in $L_i$ are presented in terms of the stationary distributions of adjacent subsystems $L_{i-1}$ and $L_{i+1}$. Unknown parameters are calculated by an iterative scheme.

In what follows, we use the following conventions. Denote the identity matrix by $\mathbf{I}$ the identity matrix of appropriate size and denote $\mathbf{I}_n$ if it is necessary to designate the size $n$ of the matrix. Similarly, $\mathbf{e}_n$ means the vector $\mathbf{e}$ of size $n$.

## 3.2 Subsystems

Subsystem $L_i$ consists of two servers $M_{i-1}$ and $M_i$ and a buffer $B_i$ between them. Since $M_i$ is a downstream server in $L_i$ and upstream server in $L_{i+1}$, denote the downstream server in $L_i$ by $M_i^d$ and the upstream server in $L_{i+1}$ by $M_i^u$, if necessary to distinguish them. Denote the states of $M_i^u$ and $M_i^d$ at time $t$ by $M_i^u(t)$ and $M_i^d(t)$, respectively. Define the the state $M_{i-1}^u(t)$ of the upstream server $M_{i-1}$ with $X_{i-1}(t)$ in $L_i$ as follows:

$$M_{i-1}^u(t) = \begin{cases} w_1(j), & \text{if } M_{i-1}(t) = j, X_{i-1}(t) = 1 \\ w_2(j), & \text{if } M_{i-1}(t) = j, X_{i-1}(t) \geq 2 \\ b_1, & \text{if } M_{i-1}(t) = b, X_{i-1}(t) = 1 \\ b_2, & \text{if } M_{i-1} = b, X_{i-1}(t) \geq 2 \\ s, & \text{if } M_{i-1}(t) = s \end{cases}$$

and the state of $M_i^d(t)$ of the downstream server $M_i$ in $L_i$

$$M_i^d(t) = \begin{cases} j, & \text{if } M_i(t) = j \\ b, & \text{if } M_i(t) = b \\ s, & \text{if } M_i(t) = s. \end{cases}$$

Let $\boldsymbol{w}_k(i) = \{w_k(1), \cdots, w_k(h_i)\}$, $k = 1, 2$, $\boldsymbol{b} = \{b_1, b_2\}$, $\boldsymbol{w}^u(i) = \boldsymbol{w}_1(i) \cup \boldsymbol{w}_2(i)$. Let

$$\{M_{i-1}^u(t) = j\} = \{M_{i-1}^u(t) \in \{w_1(j), w_2(j)\}\},$$
$$\{M_{i-1}^u(t) = b\} = \{M_{i-1}^u(t) \in \boldsymbol{b}\}.$$

Let $Z_i(t) = (X_i(t), M_{i-1}^u(t), M_i^d(t))$. The state space $\mathcal{S}_i$ of $Z_i(t)$ can be easily obtained. For example, when $X_i(t) = n$, $1 \leq n < K_i$, the set on which $Z_i(t)$ can attain the values is

$$\mathcal{S}_i(n) = \{(n, x, y) : x \in \{\boldsymbol{w}^u(i-1), s\}, y \in \{\boldsymbol{w}(i), b\}\}.$$

The stochastic process $\boldsymbol{Z}_i$ is modeled by a Markov chain on $\mathcal{S}_i$ with generator of the form

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

where $B_i^{(n)}$ is the square matrix of size $m_i(n)$ whose $((x,y),(x',y'))$-component $[B_i^{(n)}]_{(x,y),(x',y')}$ is the transition rate of $Z_i(t)$ from the state $(n,x,y)$ to the state $(n,x',y')$ without change of $X_i(t)$. Similarly, the $((x,y),(x',y'))$-component of $A_i^{(n)}$ is corresponding to the transition rate from $(n,x,y)$ to $(n+1,x',y')$ and $[C_i^{(n)}]_{(x,y),(x',y')}$ is corresponding to the the transition rate from $(n,x,y)$ to $(n-1,x',y')$.

## 3.3 Transition Rates

Note that when $X_i(t) = n$ with $1 \leq n \leq K_i - 2$, the behavior of $M_{i-1}$ does not depend on the state of $M_i$ and it depends only on the state of $X_j(t)$ and $M_j$, $j = 0, 1, \cdots, i-2$. Similarly, if $X_i(t) = 0$, then the state of $M_i$ is changed by the service completion of $M_{i-1}$, and if $X_i(t) = K_i$, then $M_{i-1}^u(t) \in \boldsymbol{b}$ and the state transition of $M_{i-1}$ occurs by a departure from $M_i$. Furthermore, if $M_{i-1}^u(t) = b^1$, then the state transition of $M_{i-1}$ occurs by an arrival from the $M_{i-2}$ as well as a departure from $M_i$. For the derivation of the rates, we adopt the approximation assumption that given $M_i(t) = x$, $X_{i-1}(t)$ and $X_i(t)$ are conditionally independent. Here, we omit details and just describe the formulae of the matrices $A_i^{(n)}$, $B_i^{(n)}$ and $C_i^{(n)}$.

**Formula of $A_i^{(n)}$.** The approximate formula for $A_i^{(n)}$ is given as follows: for $0 \leq n \leq K_i - 1$, $1 \leq i \leq N$,

$$A_i^{(n)} = A_{i-1}^u(n) \otimes A_i^d(n),$$

where $A \otimes B$ denotes the Kronecker product of the matrices $A$ and $B$. The matrices $A_i^u(n)$ and $A_i^d(n)$ are given as follows: for $0 \leq n \leq K_{i+1} - 2$, $A_0^u(n) = \boldsymbol{S}_0^0 \boldsymbol{\alpha}_0$,

$$A_i^u(n) = \begin{array}{c} \\ \boldsymbol{w}_2 \\ \boldsymbol{w}_1 \\ s \end{array} \begin{array}{c} \begin{array}{ccc} \boldsymbol{w}_2 & \boldsymbol{w}_1 & s \end{array} \\ \begin{pmatrix} q_i^u(\boldsymbol{w}_2, w_2)\boldsymbol{\alpha}_i & q_i^u(\boldsymbol{w}_1, w_1)\boldsymbol{\alpha}_i & 0 \\ 0 & 0 & \boldsymbol{S}_i^0 \\ 0 & 0 & 0 \end{pmatrix} \end{array},$$

and $A_0^u(K_1-1)=\boldsymbol{S}_0^0$,

$$A_i^u(K_{i+1}-1)=\begin{array}{c} \\ \boldsymbol{w}_2 \\ \boldsymbol{w}_1 \\ s \end{array}\begin{array}{c} \begin{array}{cc} b_2 & b_1 \end{array} \\ \begin{pmatrix} \boldsymbol{S}_i^0 & 0 \\ 0 & \boldsymbol{S}_i^0 \\ 0 & 0 \end{pmatrix} \end{array},$$

where $q_i^u(\boldsymbol{w}^2,w^1)$ and $q_i^u(\boldsymbol{w}^2,w^2)$ are the column vectors of size $h_i$ whose $j$th-entries are as follows: for $1\le j\le h_i$,

$$\begin{aligned} q_i^u(j,w_1) &= P\left(X_i(t)=2 \,\middle|\, \begin{array}{l} M_i(t)=j, \\ X_i(t)\ge 2 \end{array}\right) s_i^0(j), \\ q_i^u(j,w_2) &= s_i^0(j)-q_i^u(j,w_1), \end{aligned}$$

and for $1\le i\le N-1$,

$$A_i^d(0)=\begin{array}{c} \\ s \end{array}\begin{array}{c} \begin{array}{cc} \boldsymbol{w}_i & b \end{array} \\ \begin{pmatrix} \boldsymbol{\alpha}_i & 0 \end{pmatrix} \end{array}, \quad A_N^d(0)=\boldsymbol{\alpha}_N,$$

$A_i^d(n)=\mathbf{I}_{m_i^d}$, $1\le n\le K_i-1$, $1\le i\le N$.

**Formula of $C_i^{(n)}$.** The matrices $C_i^{(n)}$, $1\le n\le K_i$, $0\le i\le N-1$ are given by

$$C_i^{(n)}=C_{i-1}^u(n)\otimes C_i^d(n).$$

The matrices $C_i^u(n)$ and $C_i^d(n)$ are as follows: $C_0^u(K_1)=\boldsymbol{\alpha}_0$, $C_i^u(n)=\mathbf{I}_{m_i^u}$, $1\le n\le K_{i+1}-1$, and

$$C_i^u(K_{i+1})=\begin{array}{c} \\ b_2 \\ b_1 \end{array}\begin{array}{c} \begin{array}{ccc} \boldsymbol{w}_2 & \boldsymbol{w}_1 & s \end{array} \\ \begin{pmatrix} (1-p_i^u(b,s))\boldsymbol{\alpha}_i & p_i^u(b,s)\boldsymbol{\alpha}_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array},$$

where

$$p_i^u(b,s)=P(X_i(t)=2|M_i^u(t)=b,X_i(t)\ge 2)$$

and

$$C_N^d(n)=\begin{cases} \boldsymbol{S}_N^0, & n=1, \\ \boldsymbol{S}_N^0\boldsymbol{\alpha}_N, & 2\le n\le K_N, \end{cases}$$

and for $1\le i\le N-1$,

$$C_i^d(1)=\begin{array}{c} \\ \boldsymbol{w} \\ b \end{array}\begin{array}{c} \begin{array}{c} s \end{array} \\ \begin{pmatrix} \delta_i^d(\boldsymbol{w}) \\ \delta_i^d(b) \end{pmatrix} \end{array},$$

$$C_i^d(n)=\begin{array}{c} \\ \boldsymbol{w} \\ b \end{array}\begin{array}{c} \begin{array}{cc} \boldsymbol{w} & b \end{array} \\ \begin{pmatrix} \delta_i^d(\boldsymbol{w})\boldsymbol{\alpha}_i & 0 \\ \delta_i^d(b)\boldsymbol{\alpha}_i & 0 \end{pmatrix} \end{array}, \ 2\le n\le K_i,$$

where $\delta_i^d(\boldsymbol{w})$ is the column vector of size $h_i$ whose $j$th component ($1\le j\le h_i$) is

$$\begin{aligned} \delta_i^d(j) &= P(X_{i+1}(t)\le K_{i+1}-2|M_i(t)=j)s_i^0(j), \\ \delta_i^d(b) &= \sum_{y\in\mathcal{M}_{i+1}^d} P(M_{i+1}(t)=y|M_i(t)=b)\delta_{i+1}^d(y). \end{aligned}$$

**Formula of $B_i^{(n)}$.** The matrices $B_i^{(n)}$ are given as follows

$$B_i^{(n)}=\begin{cases} B_{i-1}^u-\Delta_i(0), & n=0, \\ B_{i-1}^u\oplus B_i^d-\Delta_i(n), & 2\le n\le K_i-1, \\ B_{i-1}^u(K_i)\oplus B_i^d-\Delta_i(K_i), & n=K_i, \end{cases}$$

where $A\oplus B=A\otimes\mathbf{I}+\mathbf{I}\otimes B$ denotes the Kronecker sum of the matrices $A$ and $B$ and $\Delta_i(n)$ is the diagonal matrix that makes $Q_i\mathbf{e}=0$. Denote by $S_i^*$ the square matrix of size $h_i$ whose diagonal elements are all zero and off diagonal elements are the same as those of $S_i$. The matrices $B_i^u$ and $B_i^d$ are as follows:

$$B_0^u=S_0^*, \quad B_N^d=S_N^*,$$

and for $1\le i\le N-1$,

$$B_i^u=\begin{array}{c} \\ \boldsymbol{w}_2 \\ \boldsymbol{w}_1 \\ s \end{array}\begin{array}{c} \begin{array}{ccc} \boldsymbol{w}_2 & \boldsymbol{w}_1 & s \end{array} \\ \begin{pmatrix} S_i^* & 0 & 0 \\ q_i^u(\boldsymbol{w}_1,\boldsymbol{w}_2) & S_i^* & 0 \\ & q_i^u(s,w_1)\boldsymbol{\alpha}_i & 0 \end{pmatrix} \end{array},$$

$$B_i^u(K_{i+1})=\begin{array}{c} \\ b_2 \\ b_1 \end{array}\begin{array}{c} \begin{array}{cc} b_2 & b_1 \end{array} \\ \begin{pmatrix} S_i^* & 0 \\ q_i^u(b_1,b_2) & 0 \end{pmatrix} \end{array},$$

$$B_i^d=\begin{array}{c} \\ \boldsymbol{w} \\ b \end{array}\begin{array}{c} \begin{array}{cc} \boldsymbol{w} & b \end{array} \\ \begin{pmatrix} S_i^* & q_i^d(\boldsymbol{w},b) \\ 0 & 0 \end{pmatrix} \end{array}.$$

The $q_i^u(\boldsymbol{w}_1,\boldsymbol{w}_2)$ is the diagonal matrix of size $h_i$ whose $j$th diagonal element is

$$[q_i^u(\boldsymbol{w}_1,\boldsymbol{w}_2)]_j = \sum_{k=1}^{h_{i-1}} P\left(M_{i-1}(t)=k \,\middle|\, \begin{array}{l} M_i(t)=j \\ X_i(t)=1 \end{array}\right) s_{i-1}^0(k)$$

and $q_i^d(\boldsymbol{w},b)=(q_i^d(j,b),j=1,\cdots,h_i)$ is the column vector os size $h_i$ with

$$\begin{aligned} q_i^u(b_1,b_1) &= \sum_{k=1}^{h_{i-1}} P(M_{i-1}(t)=k|M_i(t)=b^1)s_{i-1}^0(k), \\ q_i^u(s,w_1) &= \sum_{k=1}^{h_{i-1}} P(M_{i-1}(t)=k|M_i(t)=s)s_{i-1}^0(k), \\ q_i^d(j,b) &= P(X_{i+1}(t)=K_{i+1}-1|M_i(t)=j)s_i^0(j). \end{aligned}$$

## 3.4 Approximation of Parameters

Now we assume that the system is in stationary state and let $\boldsymbol{\pi}_i$ be the stationary distribution of $Q_i$. We can express the parameters $q_i^u(x,x')$, $p_i^u(b,x)$ for $A_i^u(n)$, $B_i^u$, $C_i^u(n)$ of the subsystem $L_{i+1}$ and $\delta_{i-1}^d(y)$,

$q^d_{i-1}(w,b)$ of the subsystem $L_{i-1}$ in terms of $\boldsymbol{\pi}_i$ by using the approximation

$$\begin{aligned} & P(X_i(t)=n, M_{i-1}(t)=j, M_i(t)=y) \\ = \; & \pi_i(n, w_1(j), y) + \pi_i(n, w_2(j), y), \\ & P(X_i(t)=K_i, M_{i-1}(t)=b, M_i(t)=y) \\ = \; & \pi_i(K_i, b_1, y) + \pi_i(K_i, b_2, y). \end{aligned}$$

For example,

$$\begin{aligned} q^u_i(j, w_1) &= \frac{p^d_i(2,j)s^0_i(j)}{p^d_i(w_2(j))}, \\ q^d_{i-1}(j,b) &= \frac{p^u_{i-1}(K_i-1,j)s^0_{i-1}(j)}{p^u_{i-1}(j)}, \end{aligned}$$

where

$$\begin{aligned} & p^d_i(n,y) = P(X_i(t)=n, M^d_i(t)=j), \\ & p^d_i(w_2(j)) = \sum_{n=2}^{K_i} p^d_i(n,j), \\ & p^u_{i-1}(j) = P(M^u_{i-1}(t)=j), \\ & p^u_{i-1}(n,j) = P(X_i(t)=n, M^u_{i-1}(t)=j). \end{aligned}$$

*Performance measures.* Once the stationary distribution $\boldsymbol{\pi}_i$ of $\boldsymbol{Z}_i$ is obtained, the performance measures such as the throughput $(E)$ and mean number $\bar{B}_i$ of customers in $B_i$ can be obtained as follows:

$$\begin{aligned} E &= \sum_{j=1}^{h_N} P(M_N(t)=j) s^0_N(j), \\ \bar{B}_i &= \sum_{n=1}^{K_i-1} (n-1)\boldsymbol{\pi}_i(n)\mathbf{e} + (K_i-2)\boldsymbol{\pi}_i(K_i)\mathbf{e}. \end{aligned}$$

# 4 ALGORITHM

In this section, an iterative algorithm for solving the proposed decomposition equations is presented.

**Initial Step:** Initially the upstream servers are assumed to be never starved and $A^u_i(n)$, $B^u_i(n)$ and $C^u_i(n)$ are as follows: for $0 \le i \le N-1$,

$$\begin{aligned} A^u_i(n) &= \begin{cases} \boldsymbol{S}^0_i \boldsymbol{\alpha}_i, & 0 \le n \le K_{i+1}-2, \\ \boldsymbol{S}^0_i, & n = K_{i+1}-1, \end{cases} \\ B^u_i &= \boldsymbol{S}^*_i, \\ C^u_i(n) &= \begin{cases} \mathbf{I}, & 1 \le n \le K_{i+1}-1, \\ \boldsymbol{\alpha}_i, & n = K_{i+1}. \end{cases} \end{aligned}$$

Compute the stationary distribution $\boldsymbol{\pi}_N$ and determine the parameters for $A^d_{N-1}(n)$, $B^d_{N-1}$ and $C^d_{N-1}(n)$

**Iteration Step:**

**Step 1.** Backward iteration. For $i = N-1, N-2, \cdots, 1$, compute the stationary distribution $\boldsymbol{\pi}_i$. If $i \ge 2$, then update $A^d_{i-1}(n)$, $B^d_{i-1}$ and $C^d_{i-1}(n)$, otherwise $(i=1)$ update $A^u_i(n)$, $B^u_i$ and $C^u_i(n)$.

**Step 2.** Forward iteration. For $i = 2, 3, \cdots, N$, compute the stationary distribution $\boldsymbol{\pi}_i$. If $i \le N-1$, then update $A^u_i(n)$, $B^u_i$ and $C^u_i(n)$, otherwise $(i = N)$ GO TO next step.

**Step 3.** Calculate throughput and check the tolerance. Once the stationary distribution $\boldsymbol{\pi}_N$ of $\boldsymbol{Z}_N$ is obtained, compute the throughput and check the stopping criterion

$$TOL = |E^{(m)} - E^{(m-1)}| < \varepsilon, \qquad (1)$$

where $E^{(m)}$ is the throughput obtained in the $m$th iteration and $\varepsilon > 0$ is the tolerance predetermined. If the stopping criterion is not satisfied, update $A^d_{N-1}(n)$, $B^d_{N-1}$ and $C^d_{N-1}(n)$ and GO TO Step 1 and repeat the backward and forward iteration until the stopping criterion is satisfied.

The stationary distribution $\boldsymbol{\pi}_i$ can be computed by well known matrix analytic method that uses the inversions of matrices whose sizes are the same as those of the block matrices $B_i(n)$ in $Q_i$, see e.g. Latouche and Ramaswami (1999), Shin and Moon (2017). The complexity of the algorithm in an iteration is $O(Nm_i^3)$, where $m_i$ is the maximal size of the matrices $B_i(n)$.

# 5 NUMERICAL RESULTS

The effectiveness of the method is investigated numerically in this section. Approximations (App) are compared with the simulations (Sim). Simulation models for the systems in the tables are developed with ARENA. Simulation run time is set to 550,000 unit times including 50,000 unit times of warm-up period. Twenty replications are conducted for each case and the maximum half width of 95% confidence intervals (c.i.) is less than 0.001 and confidence intervals are omitted in the following tables. Tolerance $\varepsilon = 10^{-5}$ is used for stopping criterion (1). The deviation $(\mathcal{D})$ between approximation and simulation is calculated by

$$\mathcal{D}(\%) = \frac{\text{App} - \text{Sim}}{\text{Sim}} \times 100,$$

where Sim and App are the simulation results and approximation results for throughput.

We consider the systems with 5 servers and 10 servers where the service times are identical with common means 1.0 and the buffer size between servers are identical. The buffer size $c_i$ between

Table 2: Throughput for tandem queues with 5 servers.

| $c_i$ | 1 | 5 | 10 |
|---|---|---|---|
| | Sim | Sim | Sim |
| $C_s^2$ | App($\mathcal{D}(\%)$) | App($\mathcal{D}(\%)$) | App($\mathcal{D}(\%)$) |
| 0.25 | 0.805 | 0.929 | 0.961 |
| | 0.790(-1.8) | 0.927(-0.2) | 0.961(-0.0) |
| 0.5 | 0.712 | 0.875 | 0.927 |
| | 0.696(-2.3) | 0.871(-0.5) | 0.926(-0.1) |
| 1.0 | 0.608 | 0.793 | 0.869 |
| | 0.594(-2.3) | 0.786(-0.9) | 0.867(-0.2) |
| 2.0 | 0.513 | 0.690 | 0.784 |
| | 0.512(-0.2) | 0.687(-0.4) | 0.785(+0.2) |
| 5.0 | 0.416 | 0.553 | 0.646 |
| | 0.419(+0.7) | 0.552(-0.2) | 0.650(+0.7) |

Table 3: Throughput for tandem queues with 10 servers.

| $c_i$ | 1 | 5 | 10 |
|---|---|---|---|
| | Sim | Sim | Sim |
| $C_s^2$ | App($\mathcal{D}(\%)$) | App($\mathcal{D}(\%)$) | App($\mathcal{D}(\%)$) |
| 0.25 | 0.776 | 0.917 | 0.954 |
| | 0.755(-2.8) | 0.917(-0.1) | 0.955(+0.2) |
| 0.5 | 0.673 | 0.855 | 0.915 |
| | 0.649(-3.7) | 0.851(-0.5) | 0.916(+0.1) |
| 1.0 | 0.560 | 0.763 | 0.849 |
| | 0.537(-4.1) | 0.752(-1.4) | 0.848(-0.1) |
| 2.0 | 0.449 | 0.643 | 0.749 |
| | 0.445(-0.9) | 0.634(-1.4) | 0.750(+0.1) |
| 5.0 | 0.334 | 0.487 | 0.591 |
| | 0.338(+1.2) | 0.474(-2.7) | 0.589(-0.4) |

servers are chosen as $c_i = 1,5,10$. We use the Erlang distribution of order $k$ $(E_k)$ for $C_s^2 = \frac{1}{k} < 1$, exponential distribution $(Exp)$ for $C_s^2 = 1$ and hyperexponential distribution of order 2 $(H_2)$ with balanced mean for $C_s^2 > 1$. The throughput for the system with 5 servers and 10 servers are presented in Tables 2 and 3. The deviations of approximations from simulation in Tables 2 and 3 are less than 1% for short length system $(N^* = 5)$ and are less than 5% even for moderate size of system with small buffer size $(N^* = 10, c_i = 1)$ which may be acceptable in practicable situation. Furthermore, the accuracy of approximations increases as buffer size increases and the deviations are less than 1% for $c_i = 10$. The throughput for the systems with 10 identical servers of Lognormal (LN) service times (Sim) and PH-service times (App) is listed in Table 4, where the deviation $(\mathcal{D})$ is the relative deviation between LN service system and PH service system with respect to LN service system. Table 4 shows that the deviation increases as the variation $(C_s^2)$ of service time increases or the buffer size $(c_i)$ decreases which can be expected from Section 2. It can be seen from Table 4 that the approximation performs well for the system in which the coefficient of variation of service time is not large and buffer size is not so small.

The current algorithm was run on a laptop computer at 2.70 GHz with 16.0 GB RAM using Mathematica®11. The maximum CPU time for the results in Table 3 is 1.25 seconds with 15 iterations for $E_4$ service time and others are less than 0.6 seconds.

Table 4: Throughput for the systems with 10 servers of Lognormal (LN) service times and PH-service times.

| $c_i$ | 1 | 5 | 10 |
|---|---|---|---|
| | LN | LN | LN |
| $C_s^2$ | PH ($\mathcal{D}(\%)$) | PH($\mathcal{D}(\%)$) | PH($\mathcal{D}(\%)$) |
| 0.25 | 0.770 | 0.916 | 0.953 |
| | 0.755(-2.0) | 0.917(+0.1) | 0.955(+0.2) |
| 0.5 | 0.666 | 0.851 | 0.912 |
| | 0.649(-2.8) | 0.851(+0.0) | 0.916(+0.4) |
| 1.0 | 0.558 | 0.757 | 0.844 |
| | 0.537(-3.9) | 0.753(-0.6) | 0.848(+0.4) |
| 2.0 | 0.459 | 0.646 | 0.748 |
| | 0.445(-3.1) | 0.634(-1.9) | 0.750(+0.2) |
| 5.0 | 0.355 | 0.507 | 0.604 |
| | 0.338(-5.1) | 0.474(-6.9) | 0.589(-2.6) |

# 6 CONCLUSIONS

In this paper, an approximate analysis for tandem queues with finite buffer has been presented. The service time is fitted with a phase type (PH) distribution by matching the first two moments of service times and the system with PH-service times is approximated based on the decomposition method. To reflect the dependence between consecutive stages, the states of the servers in subsystems are indicated by the state of the number of customers in upstream subsystem as well as the states of the server (blocking, starvation, working), and the transitions among the states are considered. Numerical experiments indicated that the method works reasonably.

We have used the Erlang distribution and hyperexponential distribution for fitting the first two moments of service time. There are several ways to use PH distribution matching the first two or three moments of a nonnegative random variable, see e.g. Bobbio et al. (2005), Osogami and Harchol-Balter (2006), Tijms (2003) and references therein. However, it is not easy to choose the best one among the candidates of PH distributions. The choice may depend on the model and it requires preliminary experiment.

# ACKNOWLEDGEMENTS

The first author and the second author* were supported by Basic Research Program through the

*Corresponding author

National Research Foundation of Korea (NRF) funded by the Ministry of Education, Grant Numbers NRF-2016R1D1A1A09917954 and NRF-2018R1D1A1A09083352, respectively.

## REFERENCES

Bierbooms, R., Adan, I.J.B.F., van Vuuren, M. (2011). Approximate analysis of single-server tandem queues with finite buffers. Annals of Operations Research, DOI 10.1007/s10479-011-1021-1.

Bobbio, A., Horvath, A., Scarpa, M., Telek, M. (2005). Matching three moments with minimal acyclic phase-type distributions, Stochastic Models 21, 303-326.

Buzzacott, J. A., Shanthikumar, J. G. (1993). *Stochastic Models of Manufacturing Systems*, Prentice-Hall.

Colledani, M., Tolio, T. (2011). Performance evaluation of transfer lines with general repair times and multiple failure modes, Annals of Operations Research 182, 31-65.

Dallery, Y., Gershwin, B. (1992). Manufacturing flow line systems: a review of models and analytical results. Queueing Systems 12, 3-94.

Gershwin, S. B. (1987). An effficient decomposition algorithm for the approximate evaluation of tandem queues with finite sotrage space and blocking, Operations Research 35, 291-305.

Gershwin, S. B. (1994). *Manufacturing systems engineering*. Prentice-Hall, Englewood Cliffs.

Helber, S. (2000). Approximate analysis of unreliable transfer lines with splits in the flow of material, Annals of Operations Research 93, 217-243.

Latouche, G., Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Siam.

Neuts, M. F., (1981). *Matrix Geometric Solutions in Stochastic Models*. Dover Publishing Co., New York.

Osogami, T., Harchol-Balter, M. (2006). Closed form solutions for mapping general distributions to quasi-minimal PH distributions, Performnce Evaluation 63, 524-552.

Papadopoulos, H. T., Heavey, C. (1996). Queueing theory in manufacturing systems analysis and design: a classification of models for production and transfer lines, European Journal of Operational Research 92, 1-27.

Shin, Y. W., Moon, D. H. (2017). Approximation of discrete time tandem queueing networks with unreliable servers and blocking, Journal of Industrial and Management Optimization 13(2), 901-916.

Shin, Y. W., Moon, D. H. (2018). Approximation of discrete time tandem queueing networks with unreliable servers and blocking, Performance Evaluation 120, 49-74.

Templemeier, H., Bürger, M. (2001). Performance evaluation of unbalanced flow lines with general distributed processing times, failures and imperfect production, IIE Transactions 33, 293-302.

Tijms, H. (2003). *A First Course in Stochastic Models*, Wiley.