Disposition Strategies for Open Queueing Networks with Different Service Rates

Yu-Li Tsai*, Member IAENG, Daichi Yanagisawa, and Katsuhiro Nishinari

***Abstract*—In this paper, we consider a popular kind of open queueing networks consisting of service stations arranged in series configuration with different service rate for each station. Poisson arrivals and exponential service times are assumed. We apply matrix-geometric method to evaluate steady-state probabilities and define performance measures, such as mean number in the system, mean waiting time in the system and blocking probability. Exact formulae of stability conditions are derived. Disposition strategies of service rates for each service station are suggested in order to increase working efficiency of this queueing system.**

***Keywords*—Performance Analysis, Matrix-geometric method, Disposition Strategy, Sensitivity Analysis**

I. Introduction

Increasing operational efficiency of modern automated production systems is a major for reducing operating costs in business activities. According to Road Bureau, Ministry of Land, Infrastructure, Transport and Tourism, Japan [1], the cost due to inefficient operational procedures is approximate twelve trillion Yen in Japan every year. If the circumstances of the inefficient operational procedures can be improved by considering relatively better allocation of the resources, it would make us save huge wastes of production costs. In this work, we focus on decreasing the cost of production by suggesting disposition strategies for manufacturing industries utilizing automated production systems.

Open queueing networks with no immediate waiting space are popular in real industrial applications (e.g. production line systems and I/O devices). The automated production line systems are applied widely in automotive industries. In this study, a popular open queueing networks with no buffers between each service station is investigated. This kind of queuing system is very common in modern automotive production line and in manufacturing processes in semiconductor industries. We further suggest counterintuitive disposition strategies based on the simulation results. Intuitively, the disposition strategies for keeping high operational performances (i.e. reducing mean waiting time of the system) for this kind of system consisting of the arbitrary number of service stations are all the same. However, we discover that the disposition strategies for increasing the operational efficiency of the system depending on the number of service stations are different. Especially, automobile companies can benefit from our results to increase their operational efficiency in the production line and save related time cost.

The literature on the series configuration queueing system with blocking phenomena can be traced back to Hunt [2]. He studied four particular cases of service facilities in series including infinite storage space between stages, no storage space between stages, finite storage space between stages, and the case of the unpaced belt-production line. Avi-Itzhak et al. [3] investigated a queueing system consisting of a sequence of two service stations with infinite queue allowable before the first station and no queue allowable between the stations. They obtained the moment generating functions of the steady-state queueing times and the generating functions of the steady numbers of customers in the various parts of the system. Avi-Itzhak et al. [4] studied a queueing system with sequence stations following an ordered service type. They assumed the arrival process is arbitrary and the time to serve each customers at the working stations is regular. Altiok [5] presented an approximate method for the analysis of open networks of queues in tandem and with blocking phenomena. He evaluated the steady-state probabilities of the number of customers at each station based on a specific method of decomposition where the total network is broken down into queues. Langaris et al. [6] provided a method to analyze the waiting time of a two-stage queueing system with blocking phenomena. They further considered the separation of the concepts between effective service time and the blocked time in the first service station. Papadopoulos et al. [7] developed an algorithm to model characteristics of production lines with no intermediate buffers. The marginal probability distribution of the number of units in each machine, the mean queue length and the throughput of the system can be obtained by their method. Avi-Itzhaket al. [8] assumed the just-in-time input for a queueing system with no buffers between servers under the communication and the blocking schemes. Akyildiz et al. [9] derived the exact equilibrium state probability distributions for two-station queueing networks with blocking-after-service mechanism. Avi-Itzhaket al. [10] generalized a queueing system under k-stage blocking. They discovered a result that for $k > 1$, the waiting times are not order-insensitive while the G/D/1 equivalence is maintained.

Mathematical analysis and related applications of matrix-geometric method was systematically studied by Neuts [11]. Gomez-Corral [12] applied a general theory on

Manuscript received June 19, 2016; This work was supported in part by the JSPS KAKENHI Grant Numbers 25287026 and 15K17583.

Y. L. Tsai*, corresponding author, is with the Department of Advanced Interdisciplinary Studies, School of Engineering, The University of Tokyo, CO 153-8904 Japan (phone: +81-3-5452-5022; e-mail: tokyotsai@gmail.com).

D. Yanagisawa is with the Research Center for Advanced Science and Technology, The University of Tokyo. CO 153-8904 Japan (e-mail: tDaichi@mail.ecc.u-tokyo.ac.jp)

K. Nishinari is with the Research Center for Advanced Science and Technology, The University of Tokyo. CO 153-8904 Japan (e-mail: tknishi@mail.ecc.u-tokyo.ac.jp)

quasi-birth-and-death processes to study a special kind of queueing system with blocking and repeated attempts. Gomez-Corral [13] investigated a two-stage tandem queue with blocking under the presence of a secondary flow of disasters. He determined the stationary distribution at departure epochs by using spectral analysis and calculated the stationary distribution at an arbitrary time. Gomez-Corral [14] studied queueing networks with blocking under the assumption of input units follow Markovian Arrival Process and applied the general theory on Markov renewal processes of M/G/1-type in their analysis. Gomez-Corral et al. [15] considered a two-stage tandem G-queue with blocking, service requirements of phase type and arrivals of units and of signals. They further investigated the influence of several flows of signals on the performance evaluation of the queueing model through various probabilistic descriptors. Gomez-Corral et al. [16] studied the influence of the dependence between units and signals on the performance evaluation of the continuous-time Markov chain describing the state of the network at arbitrary times. Bierbooms et al. [17] developed approximate methods for fluid flow production lines with multi-server workstations and finite buffers. Their method is suitable for the estimations of characteristics of longer production lines. Bierbooms et al. [18] proposed an approximation method to determine the throughput and mean sojourn time of single server tandem queues with general service times and finite buffers by decomposition method. Zhou et al. [19] studied a two-stage tandem queueing network with Markovian arrival process inputs and buffer sharing. They discovered that the buffer sharing policy is more flexible when the inputs have large variant and are correlated. Hillier [20] considered the optimal design of unpaced assembly lines. He analyzed the joint optimization of both the allocation of workload and the allocation of buffer spaces simultaneously when the objective is to maximize the revenue from throughput minus the cost of work-in-process inventory. Sakuma et al. [21] proposed Whitt's approximation to obtain the stationary distribution of an assembly-like queueing system with generally distributed time-constraint. Shin et al. [22] developed an approximation method for throughput in tandem queues with multiple independent reliable servers at each stage and finite buffers between service stations. Hudson et al. [23] gave complete reviews for the topics about unbalanced unpaced serial production lines. Several unanswered questions about the performance of assembly line are described in this work. Sani and Daman [25] studied a M/G/2 queueing system with an exponential server and a general server under a controlled queue discipline. The steady state distribution for the number of customers in the system, mean waiting time, mean queue length and blocking probability for the queueing system are derived. Ramasamy et al. [26] presented the steady state analysis of a heterogeneous server queueing system, Geo/G/2. Services containing discrete in nature can be applied through their analysis in many areas of communication, telecommunications, business and computer systems. Tsai et al. [27] discussed series configuration queueing systems with four service stations. They proposed general disposition strategies of the system based on original inductions of this works. Tsai et al. [28] further extended the series configuration queueing system by considering the conditions of system breakdowns and repairs. The disposition strategies of this kind of queueing system are suggested according to their theoretical and numerical investigations.

We further cite surveys and bibliographies in this important topic by Perros [29-30], Onvural [31], Balsamo [32] and Hall et al. [33], two major monographs by Perros [34] and Balsamo et al. [35], and other special collections from journals [36-37]. Decomposition methods applied to study tandem queueing systems can be referred to Hillier et al. [38] and Perros et al. [39].

This study is the original works to propose the idea of disposition strategies for open queueing networks with different service rates in order to increase the operational efficiency of this popular kind of queueing systems. Moreover, we further discover that disposition strategies depend on the number of service stations of the system through numerical simulations. The exact results of maximum utilization of the system consisting of two and three service stations are also successfully explained by numerical results. Other theoretical results reveal that if one of the service stations breaks down without any repair processes, the utilization of the system become useless. We expect that our results can be applied for increasing the operational efficiency of production line systems in automobile industry.

Our major contributions are following:

- Methodological.

We analyze open queueing networks with blocking phenomena, in particular:

1. Constructing steady-state equations and structured generator matrix of the queueing system with two and three service stations.
2. Giving exact formulae of stability conditions for the system with two and three service stations.
3. Solving the steady-state probabilities with different service rates.
4. Proposing disposition strategies for the system with different service rates of each service station working in high performance ways through simulations.
5. Observing that the maximum utilization of the system decreases as the number of service stations of the system increases.
6. Exact theoretical values of the maximum utilization of the system are clearly explained corresponding to numerical results.
7. Suggesting that the management of companies should prepare repair processes for the system if there are possibilities of happening breakdowns of service stations for the series configuration queueing system.

- Practical.

We model and analyze the control of service rates for each service station of the system from the viewpoint of practitioners.

1. Our model capture important performance measures, such as blocking probability of each service station and mean waiting in the system which provide theoretical predictions of the characteristics of the system.
2. They yield insights for controlling finite resources to increase the operational efficiency in real applications.
3. Companies in the automobile industry and other industries using similar assembly line systems can apply our results to improve operational efficiency of production line systems through internet of things (IOT) technologies and real-time data analytics.

Paper Outline: The rest of the paper is organized as follows. The formal description of the problem and the summary of notations used in our model is introduced in the beginning of next section.

Detailed descriptions of matrix-geometric method employed to the system with two and three service stations and major performance measures for the system are given in section 3. Numerical results and the suggested disposition strategies for the case studies of the system are presented in section 4. Finally, we conclude with discussions of our works and indicate possible directions for future research in section 5.

II. Problem formulation and notations

In our analysis, we assume a queueing system consisting of independent service stations in series configuration and operating simultaneously. The series configuration system with three service stations is depicted in Figure 1. Each customer entering to the system follows Poisson arrival process with mean arrival rate λ. The time to serve a customer in each station is exponentially distributed with mean service time $\frac{1}{\mu}$. Each customer should enter all of the service stations in order to finish complete services. A complete service is defined as after finishing all jobs in each service station, the customer leaves the queueing system. There are no intermediate waiting lines between service stations, so another main characteristic in the system is blocking. This phenomenon called blocking after service happens in the case that a customer completes the service in a service station, but another customer in the next station has not finished the service yet. The customer who completed the service is blocked by the customer who is still receiving the service located next station. An infinite capacity queue is allowed in front of the first service station. In addition, only one customer can enter each service station at a time and the service rate is independent of the number of customers. The service of the system obeys the first come first serve (FCFS) discipline.

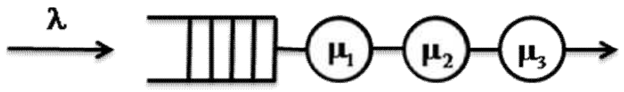

Figure 1. Series configuration queueing system with three service stations.

- **Notations**

In this section, we introduce notations used in our model framework. Mean arrival rate of Poisson arrivals is denoted as λ. We reserve the notations μ_1 and μ_2 and μ_3 for the mean service rate of the station-1 and the station-2 and the station-3, respectively. We use P_{n_1,n_2,n_3,n_4} to denote the steady-state probability P_{n_1,n_2,n_3,n_4} of n_1 customer in the station-3 and n_2 customer in the station-2 and n_3 customer in the station-1 and n_4 customer in the queue. For instance, the steady-state probability $P_{0,1_b,6}$ means that there is a customer who is blocked in the station-1, since the customer in the station-2 is still receiving the service. There is no customer in the station-3 and 6 customers waiting in the queue. Similarly, for the system consisting of two service stations, the notation P_{n_1,n_2,n_3} means the steady-state probability P_{n_1,n_2,n_3} of n_1 customer in the station-2 and n_2 customer in the station-1 and n_3 customer in the queue.

III. Modeling Framework

- **Matrix-Geometric Method**

We denote $\mathbf{P} = [\mathbf{P_0}, \mathbf{P_1}, \mathbf{P_2}, ...]$ as steady-state probability vector corresponding to the transition matrix Q. Note that the steady-state probability vector comprises steady-state probabilities of the quasi-death-birth process. The detailed compositions of the sub-matrices of the transition matrix Q for the system with two and three service stations are given in Appendix. The equilibrium equation of the quasi-birth-death process can be described as $\mathbf{PQ} = \mathbf{0}$, while the normalization condition of the steady-state probability is $\mathbf{P1} = 1$. Then, the global balance equations of the quasi-birth-death process can be written as

$$\mathbf{P_0}B_{0,0} + \mathbf{P_1}B_{1,0} = \mathbf{0}, \quad (1)$$

$$\mathbf{P_0}B_{0,1} + \mathbf{P_1}A_1 + \mathbf{P}_2A_2 = \mathbf{0}, \quad (2)$$

$$\mathbf{P}_i A_0 + \mathbf{P}_{i+1}A_1 + \mathbf{P}_{i+2}A_2 = \mathbf{0}, \qquad i \geq 1. \quad (3)$$

There exist a rate matrix R, and the following recurrence relation can be constructed

$$\mathbf{P}_i = \mathbf{P}_{i-1}R = \mathbf{P}_1 R^{i-1}, \qquad i \geq 1. \quad (4)$$

The unknown rate matrix R can be obtained by substituting (4) into (3), and simply to matrix quadratic equation

$$A_0 + RA_1 + R^2 A_2 = 0. \quad (5)$$

The simplified equations of (1) and (2) can be represented as

$$\mathbf{P_0}B_{0,0} + \mathbf{P_1}B_{1,0} = \mathbf{0}, \quad (6)$$

$$\mathbf{P_0}B_{0,1} + \mathbf{P_1}(A_1 + RA_2) = \mathbf{0}. \quad (7)$$

According to Bloch et al. [24], the normalization condition equation that only involves $\mathbf{P_0}$ and $\mathbf{P_1}$ is given by

$$\mathbf{P_0}\mathbf{1} + \mathbf{P_1}(I - R)^{-1}\mathbf{1} = 1, \quad (8)$$

where I is the identity matrix with same size as the rate matrix R.

We apply an iterative method by successive substitution, described in Neuts [11] to solve the rate matrix R from (5). Taking (6), (7) and (8) into account, the steady-state probability vector of $\mathbf{P_0}$ and $\mathbf{P_1}$ can be obtained by solving following matrix equation

$$(\mathbf{P_0}, \mathbf{P_1}) \begin{pmatrix} B_{0,0} & B_{0,1}^* & \mathbf{1} \\ B_{1,0} & (A_1 + RA_2)^* & (I - R)^{-1}\mathbf{1} \end{pmatrix} = (\mathbf{0}, 1). \quad (9)$$

where $(.)^*$ indicates that the last column of the included matrix is removed to avoid linear dependency.

- **Stability Conditions**

This section provides exact formulae of stability conditions for the system consisting of two and three service stations with equivalent and different service rates. For the reasons of stability and ergodicity, the stability condition is given by Neuts [11],

$$\mathbf{P}_A A_0 \mathbf{1} < \mathbf{P}_A A_2 \mathbf{1}, \tag{10}$$

where $\mathbf{P}_A$ is the steady-state probability vector corresponding to the generator matrix A.

The conservative stable matrix is defined to be

$$A = A_0 + A_1 + A_2. \tag{11}$$

Solving the following system equations with normalization condition, we can obtain the steady-state probability $\mathbf{P}_A$.

$$\mathbf{P}_A A = \mathbf{0}, \tag{12}$$

$$\sum_{i=0}^{2} P_{A,i} = 1. \tag{13}$$

Substituting the steady-state probability into (10) and employing the content of the matrix A_0 and A_2, the exact formulae of stability conditions can be derived.

- **Exact results of stability conditions**

Theorem 1. The stability conditions of the system consisting of two service stations

(1) For $\mu_1 \neq \mu_2$,

$$\lambda < \frac{\mu_1(\mu_1\mu_2 + \mu_2^2)}{\mu_1^2 + \mu_1\mu_2 + \mu_2^2}.$$

(2) Special case: $\mu_1 = \mu_2 = \mu$

$$\lambda < \frac{2}{3}\mu.$$

Note that the exact results of the service stations with equivalent service rate show how the maximum utilization of the queueing system is decreased by removing buffers between each service station. The maximum utilization of the system with two service stations reduces to 2/3 (i.e. about 0.667) compared with traditional tandem queuing system which contains an infinite queue between service stations. We infer this value of maximum utilization comes from the blocking probability of the station-1 in the condition of very high arrival rate is 1/3 (i.e. about 0.333). The results of the blocking probability of the station-1 of the system with two service stations can be checked in Section 3.1.

Theorem 2. The stability conditions of the system consisting of three service stations

(1) For $\mu_1 \neq \mu_2 \neq \mu_3$

$$\lambda < \frac{N_3}{D_3},$$

where

$$N_3 = \mu_1\mu_2\mu_3(\mu_1+\mu_2)(\mu_2+\mu_3)$$
$$(\mu_1^3 + \mu_1^2\mu_2 + 3\mu_1^2\mu_3 + \mu_1\mu_2\mu_3 + 3\mu_1\mu_3^2 + \mu_2\mu_3^2 + \mu_3^3),$$

and

$$\begin{aligned} D_3 &= \mu_1^5(\mu_2^2 + \mu_2\mu_3 + \mu_3^2) \\ &+ \mu_1^4(2\mu_2^3 + 5\mu_2^2\mu_3 + 5\mu_2\mu_3^2 + 3\mu_3^3) \\ &+ \mu_1^3(\mu_2^4 + 5\mu_2^3\mu_3 + 8\mu_2^2\mu_3^2 + 7\mu_2\mu_3^3 + 3\mu_3^4) \\ &+ \mu_1^2(\mu_2^4\mu_3 + 5\mu_2^3\mu_3^2 + 8\mu_2^2\mu_3^3 + 5\mu_2\mu_3^4 + \mu_3^5) \\ &+ \mu_1(\mu_2^4\mu_3^2 + 5\mu_2^3\mu_3^3 + 5\mu_2^2\mu_3^4 + \mu_2\mu_3^5) \\ &+ (\mu_2^4\mu_3^3 + 2\mu_2^3\mu_3^4 + \mu_2^2\mu_3^5). \end{aligned}$$

(2) Special case: $\mu_1 = \mu_2 = \mu_3 = \mu$

$$\lambda < \frac{22}{39}\mu.$$

In the case of the system with three service stations, the blocking probability of the first service station in the condition of very high arrival is approximate 17/39 (i.e. about 0.436), this reflect the intuition that the theoretical value of maximum utilization of the system with three service stations is 22/39 (i.e. about 0.564). Numerical simulations about blocking probability of the station-1 of the system with three service stations in Section 3.2. confirm our inferences regarding the maximum utilization of the system with three service stations.

Next, we study the behavior of the system consisting of three service stations. We consider taking the limit of the service rate of each service station to zero, respectively

$$\lim_{\mu_1 \to 0} \frac{N_3}{D_3} = 0,\ \lim_{\mu_2 \to 0} \frac{N_3}{D_3} = 0,\ \lim_{\mu_3 \to 0} \frac{N_3}{D_3} = 0.$$

The results show that if one of the service rates of the service stations approaches to zero, the mean arrival rate also should be lowered to zero. This means that if one of the service stations is failed, the impact to the whole queueing system is fateful. Since the service rate of any service stations down to zero, the number of customers in the queue would growth rapidly and tend to diverge.

- **Performance metrics**

In this section, we define the performance metrics for the series configuration system consisting of two and three service stations. Important performance measures include mean number in the system, mean number in the queue, mean waiting time in the system, mean waiting time in the queue and blocking probability of the service stations in front of the terminal station.

Performance measures for the system consisting of two service stations are defined by

(1) Mean number of customers in the system

$$L = (P_{1,0,0} + P_{0,1,0} + P_{1,b,0}) + \sum_{n=2}^{\infty} (P_{1,b,n-1} + P_{1,1,n-2} + P_{0,1,n-1}) \cdot n. \tag{14}$$

(2) Mean number of customers in the queue

$$L_q = \sum_{n=1}^{\infty} (P_{1,b,n} + P_{1,1,n} + P_{0,1,n}) \cdot n. \tag{15}$$

(3) Mean waiting time in the system (Little's Law)

$$W = \frac{L}{\lambda}. \tag{16}$$

(4) Mean waiting time in the queue (Little's Law)

$$W_q = \frac{L_q}{\lambda}. \qquad (17)$$

(5) Blocking probability of the customer in the station-1

$$P_b = \sum_{n=0}^{\infty} P_{1,b,n}. \qquad (18)$$

Performance measures for the system consisting of three service stations are defined by

(6) Mean number of customer in the system

$$\begin{aligned} L_3 &= (P_{0,0,1,0} + P_{0,1,0,0} + P_{1,0,0,0} + P_{1,b,0,0} + P_{0,1,b,0}) \\ &+ 2(P_{0,0,1,1} + P_{0,1,1,0} + P_{1,0,1,0} + P_{1,1,0,0}) \\ &+ \sum_{n=3}^{\infty}(P_{0,0,1,n-1} + P_{0,1,1,n-2} + P_{1,0,1,n-2} + P_{1,1,1,n-3}) \cdot n \\ &+ \sum_{n=2}^{\infty}(P_{1,b,1,n-2} + P_{0,1,b,n-1} + P_{1,1,b,n-2}) \cdot n \\ &+ \sum_{n=1}^{\infty}(P_{1,b,b,n-1}) \cdot n. \end{aligned} \qquad (19)$$

(7) Mean number of customers in the queue

$$\begin{aligned} L_{q,3} &= (P_{0,0,1,1} + P_{0,1,1,1} + P_{1,0,1,1} + P_{0,1,b,1}) + 2(P_{0,0,1,2}) \\ &+ \sum_{n=3}^{\infty}(P_{0,0,1,n}) \cdot n + \sum_{n=2}^{\infty}(P_{0,1,1,n} + P_{1,0,1,n} + P_{0,1,b,n}) \cdot n \\ &+ \sum_{n=1}^{\infty}(P_{1,1,1,n} + P_{1,b,1,n} + P_{1,b,b,n} + P_{1,1,b,n}) \cdot n. \end{aligned} \qquad (20)$$

(8) Mean waiting time in the system (Little's Law)

$$W_3 = \frac{L_3}{\lambda}. \qquad (21)$$

(9) Mean waiting time in the queue (Little's Law)

$$W_{q,3} = \frac{L_{q,3}}{\lambda}. \qquad (22)$$

(10) Blocking probability of the customer in the station-1

$$P_{b,1} = \sum_{n=0}^{\infty} P_{1,b,b,n} + P_{0,1,b,n} + P_{1,1,b,n}. \qquad (23)$$

(11) Blocking probability of the customer in the station-2

$$P_{b,2} = \sum_{n=0}^{\infty} P_{1,b,b,n} + P_{1,b,0,n}. \qquad (24)$$

(12) Blocking probability of the customer in the station-1 and the station-2

$$P_{b,12} = \sum_{n=0}^{\infty} P_{1,b,b,n}. \qquad (25)$$

III. Numerical Results

In this section, we perform numerical experiments for the queueing system consisting of two and three service stations. In each case, we present performance metrics of the system with equivalent service rates (i.e. $\mu_1 = \mu_2 = \mu_3 = \mu$) and different service rates. According to the results of simulation, we will suggest better disposition strategies to increase operational efficiency for the system.

3.1. Two service stations

- **Same service rates for each service station**

We first present how mean number in the system and blocking probabilities of the station-1 change as the mean arrival rate varies λ from 0.01 to 0.66, as shown in Figure 2 and Figure 3, respectively. We observe that mean number in the system increases as λ increases. This result also verifies the stability condition we have derived in the Section 2. The trends of blocking probabilities also increase and finally approach to about 0.33.

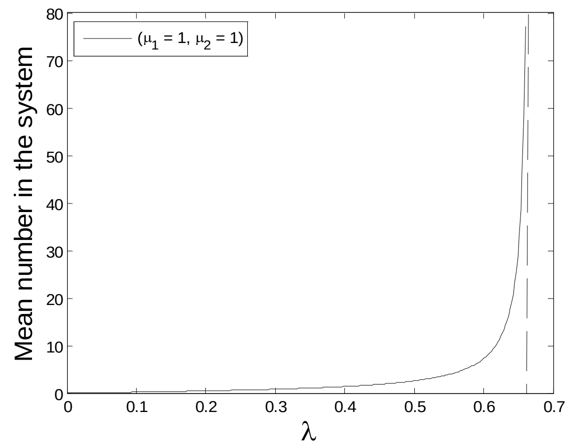

Fig 2. Mean number in the system (2 stations)

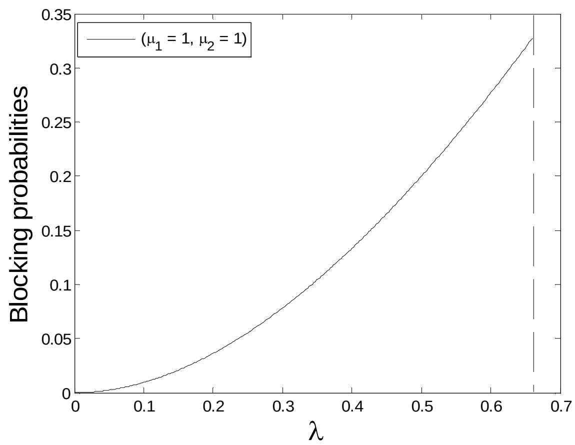

Fig 3. Blocking probability (2 stations)

- **Different service rates for each service station**

Next, we study the impact of different rates which result in significantly distinct performances for the series configuration system. We set $\mu_1 = 2, \mu_2 = 1$ and $\mu_1 = 1, \mu_2 = 2$, then vary the mean arrival rate λ from 0.01 to 0.60. It is discovered that when we set higher service rate for the station-1, the mean waiting time in the system is less than that of setting higher service rate for the station-2. Since setting higher service rate for the station-1 would advance the opportunity for the customers waiting in the queue to enter the first service station, it can reduce mean waiting time in the queue for customers, as show in Figure 4 and Figure 5, respectively. Although setting higher service rate for the station-1 may cause relatively higher blocking probability of the station-1 compared with the results of setting higher service rate for the station-2, this would not cause significant effect since the happening of blocking phenomenon in this case is still relatively low, as shown in Figure 6. We suggest setting higher service rate for the station-1 of the system with two service stations in order to maintain higher operational efficiency when the service rate of each service station is different.

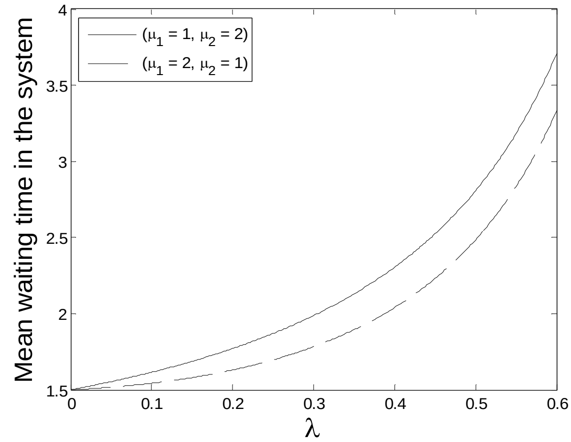

Fig 4. Mean waiting time in the system with different service rates (2 stations)

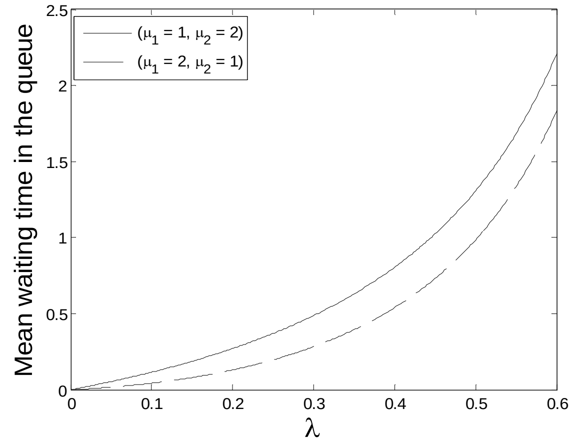

Fig 5. Mean waiting time in the queue with different service rates (2 stations)

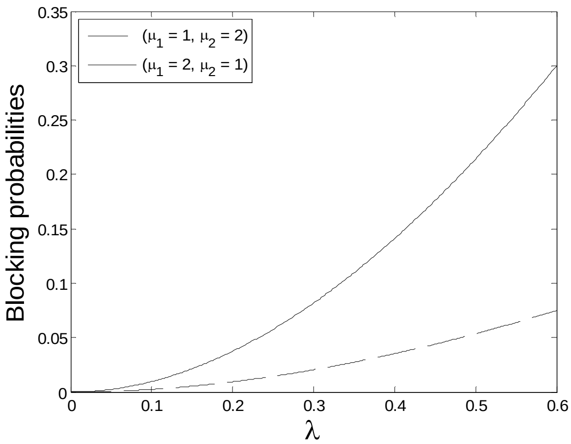

Fig 6. Blocking probability with different service rates (2 stations)

3.2. Three service stations

- **Same service rates for each service station**

The mean number in the system and blocking probability of the station-1 and the station-2 as a function of mean arrival rate of the system consisting of three service stations are shown in Figure 7 and Figure 8, respectively. The numerical results of mean number in the system are consistent with the exact results of stability conditions we derived in the section 2, which shows the upper bound of the stability condition approach to $\frac{22}{39}$ (≈ 0.564). In addition, it is noted that the blocking probability of the station-1 is higher than that of the station-2. The blocking probability of the station-1 and the station-2 happening simultaneously is relatively low in this case.

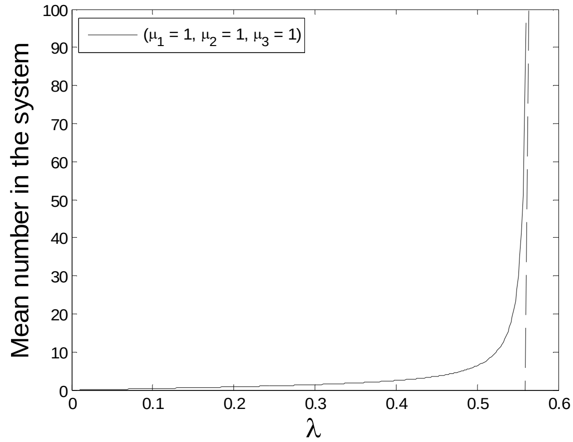

Fig 7. Mean number in the system (3 stations)

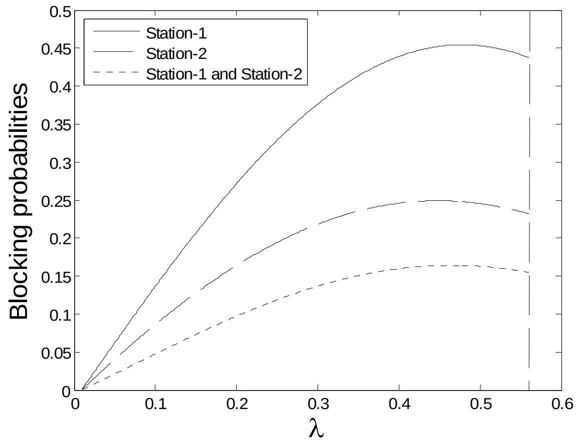

Fig 8. Blocking probability (3 stations)

- **Controlling the service rates of the two service stations**

In the case of different service rates, we assume that we are able to control the service rates of two service stations and the service rates of one service station at one time. Intuitively, it is better to set higher service rates for the service stations before the terminal service station (i.e. the last service station of the series configuration queueing system) according to the results of the system with two service stations. In the case of controlling service rates of two service stations, we set $\mu_1 = 2, \mu_2 = 2, \mu_3 = 1$ and $\mu_1 = 1, \mu_2 = 2, \mu_3 = 2$ and $\mu_1 = 2, \mu_2 = 1, \mu_3 = 2$, then vary the mean arrival rate λ from 0.01 to 0.75. The numerical results suggest that setting higher service rates for the station-2 and the station-3 result in best operational efficiency, as shown in Figure 9 and Figure 10, respectively. Since setting lower service rate for the station-1 would cause less waiting in the queue, it makes easier for the customer to enter service stations to receive their services.

It is observed that the mean waiting time in the system of the case $\mu_1 = 2, \mu_2 = 2, \mu_3 = 1$ is higher than that of the case $\mu_1 = 2, \mu_2 = 1, \mu_3 = 2$ when the mean arrival rate is lower than 0.56. This result reveals the fact that when the mean arrival rate is lower than 0.56, it takes longer time to complete services in the service stations for the case $\mu_1 = 2, \mu_2 = 2, \mu_3 = 1$, since the mean waiting time in the queue is almost the same for both cases. The setting of higher service rates for the station-1 and the station-2 would increase blocking probability of the station-1 and the station-2, so it is the major reason that the customers take longer time to receive services in the case of $\mu_1 = 2, \mu_2 = 2, \mu_3 = 1$, as shown in Figure 11, Figure 12 and Figure 13, respectively. When the mean arrival is greater than 0.56, the mean waiting time in the queue in the case of $\mu_1 = 2, \mu_2 = 1, \mu_3 = 2$ becomes relatively longer than that of the case $\mu_1 = 2, \mu_2 = 2, \mu_3 = 1$. We finally observed that the mean waiting time in the system in the case of $\mu_1 = 2, \mu_2 = 1, \mu_3 = 2$ is larger than that of the case $\mu_1 = 2, \mu_2 = 2, \mu_3 = 1$.

We suggest $\mu_1 = 1, \mu_2 = 2, \mu_3 = 2$ as the best disposition strategy, when we are able to control service rates of two service stations for the system.

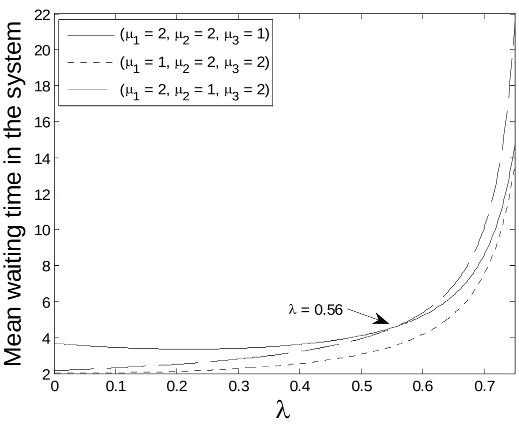

Fig 9. Mean number in the system with different service rates by controlling two service stations

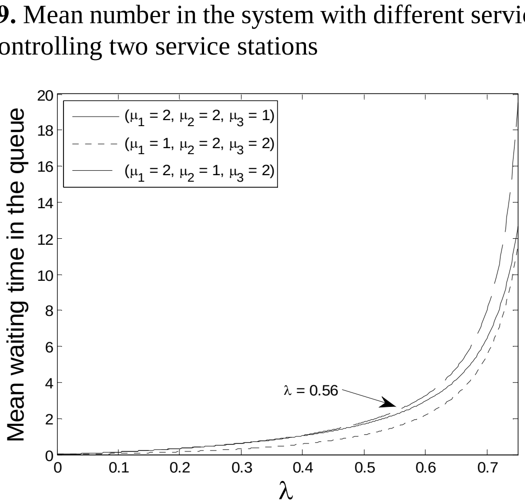

Fig 10. Mean number in the queue with different service rates by controlling two service stations

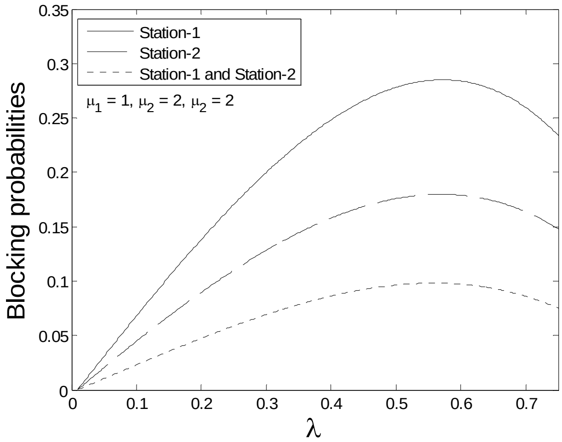

Fig 11. Blocking probability with different service rates by controlling two service stations, $\mu_1 = 1,\ \mu_2 = 2,\ \mu_3 = 2$

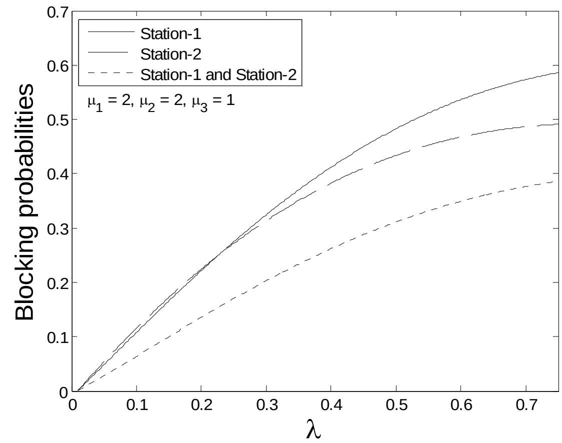

Fig 12. Blocking probability with different service rates by controlling two service stations, $\mu_1 = 2,\ \mu_2 = 2,\ \mu_3 = 1$

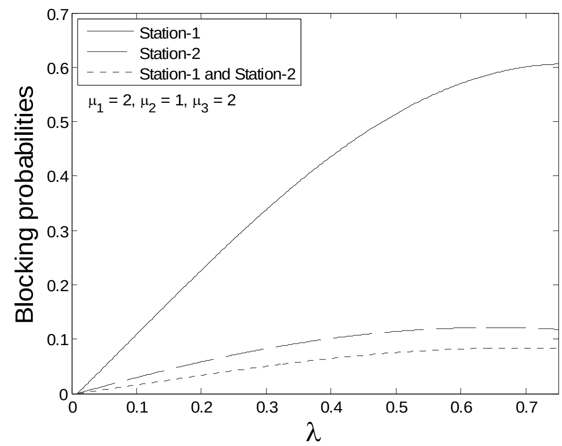

Fig 13. Blocking probability with different service rates by controlling two service stations, $\mu_1 = 2,\ \mu_2 = 1,\ \mu_3 = 2$

- **Controlling the service rates of the one service station**

Finally, we study the case of controlling service rate of one service station, we set $\mu_1 = 2, \mu_2 = 1, \mu_3 = 1$ and $\mu_1 = 1, \mu_2 = 2, \mu_3 = 1$ and $\mu_1 = 1, \mu_2 = 1, \mu_3 = 2$, then vary the mean arrival rate λ from 0.01 to 0.6. The plots are presented in Figure 14 and Figure 15, which shows that in the case of $\mu_1 = 2, \mu_2 = 1, \mu_3 = 1$, the mean waiting time is the greatest compared with other two cases. In this disposition strategy, the customers in the queue are difficult to enter the service stations, because of the high blocking probability of the station-1.

It is noted that the mean waiting time in the system of the case $\mu_1 = 1, \mu_2 = 1, \mu_3 = 2$ is lower than that of the case $\mu_1 = 1, \mu_2 = 2, \mu_3 = 1$ when the mean arrival rate is lower than 0.46. Since the mean waiting time in the queue is almost the same for both cases, we discover that the case $\mu_1 = 1, \mu_2 = 2, \mu_3 = 1$ would result in longer time to complete the services in the service stations, because of the increasing blocking probabilities of the station-1 and the station-2, as shown in Figure 16, Figure 17 and Figure 18, respectively. The waiting time in the queue in the case of $\mu_1 = 1, \mu_2 = 1, \mu_3 = 2$ becomes relatively longer than that of the case $\mu_1 = 1, \mu_2 = 2, \mu_3 = 1$, when the mean arrival is greater than 0.46. It is observed that the mean waiting time in the system in the case of $\mu_1 = 1, \mu_2 = 1, \mu_3 = 2$ is larger than that of the case $\mu_1 = 1, \mu_2 = 2, \mu_3 = 1$. The setting of lower service rates in the station-1 and the station-2 makes customers take longer waiting time in the queue.

It is suggested that set $\mu_1 = 1, \mu_2 = 1, \mu_3 = 2$ as the best disposition strategy when the mean arrival rate is lower than 0.46. Conversely, the case of $\mu_1 = 1, \mu_2 = 2, \mu_3 = 1$ is a relatively better disposition strategy when the mean arrival rate becomes larger than 0.46 in the case that we can control only one of the service rates for the system consisting of three service stations.

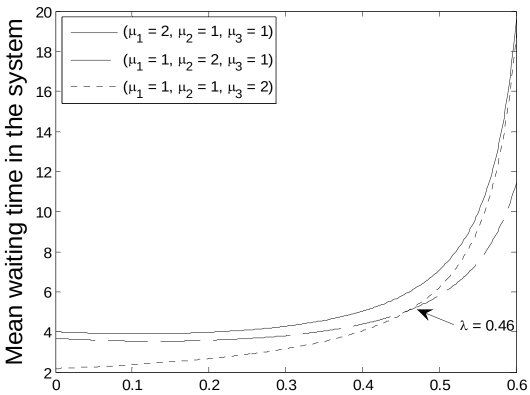

Fig 14. Mean waiting time in the system by controlling one service station

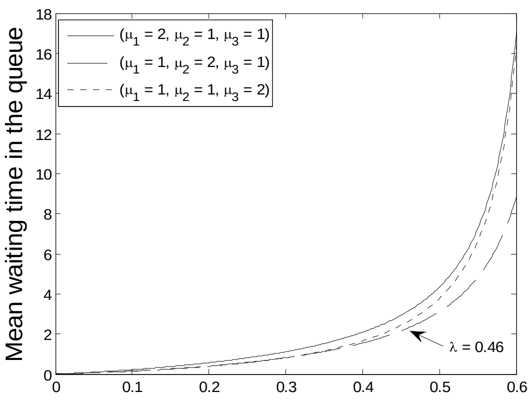

Fig 15. Mean waiting time in the queue by controlling one service station

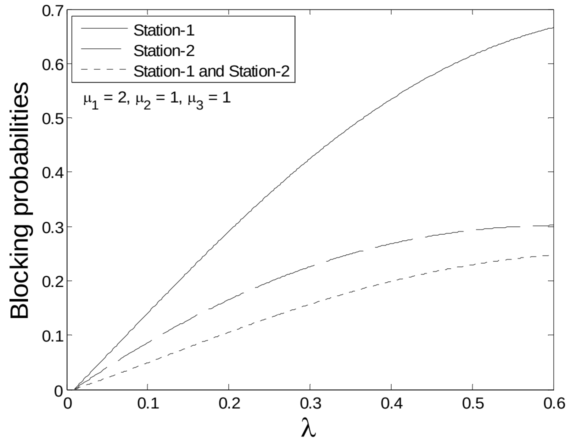

Fig 16. Blocking probability with different service rates by controlling one service station, $\mu_1 = 2,\ \mu_2 = 1,\ \mu_3 = 1$

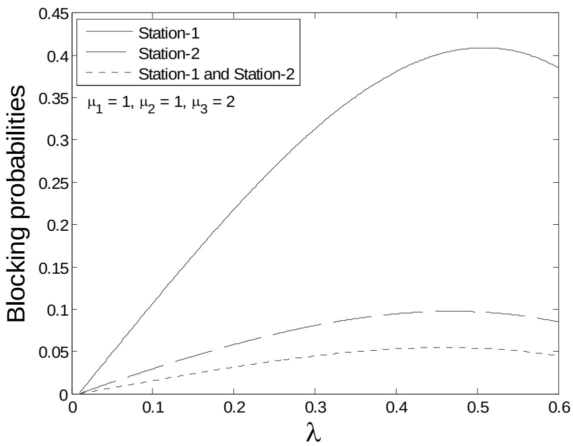

Fig 17. Blocking probability with different service rates by controlling one service station, $\mu_1 = 1,\ \mu_2 = 1,\ \mu_3 = 2$

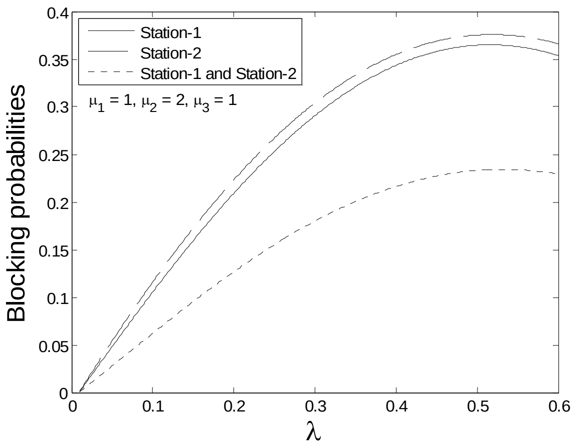

Fig 18. Blocking probability with different service rates by controlling one service station, $\mu_1 = 1,\ \mu_2 = 2,\ \mu_3 = 1$

IV. CONCLUSIONS

We have shown that matrix-geometric method is useful to evaluate steady-state probability of complex quasi-birth-death processes. Therefore, we can further understand characteristics of the series configuration queueing system by defining important performance measures such as mean number in the system, mean number in the queue, blocking probabilities, mean waiting time in the system and mean waiting time in the queue. We have further suggested disposition strategies for the system consisting of two and three service stations through the numerical results of these performance measures.

The numerical results of the system with two service stations show that it is better to set higher service rate for the station-1 in order to increase the operational efficiency of the system. Intuitively, this disposition strategy (i.e. arrange higher service rate for the service stations in front of the terminal station) can be applied to the series configuration queueing system consisting of more than 2 service stations. Surprisingly, the simulations presented the opposite results on the intuitions. We have also given better disposition strategy for the system consisting of three service stations according our case studies. If we are able to control two service rates of the service stations, the case $\mu_1 = 1, \mu_2 = 2, \mu_3 = 2$ is the best disposition strategy. On the other hand, when we just can control only one of the service rate of the service stations, the case $\mu_1 = 1, \mu_2 = 1, \mu_3 = 2$ and the case $\mu_1 = 1, \mu_2 = 2, \mu_3 = 1$ would be better depending on the conditions of mean arrival rate.

The simulation results show that the disposition strategies for this kind of queueing system with the different number of service stations are distinct (i.e. the even number of service stations vs. the odd number of service stations). We should arrange higher service rate for the service stations near the entrance of the system with the even number of service stations. On the other hand, it is suggested that setting higher service rate for the service stations located before the terminal station of the system with the odd number of service stations. This proposition is valuable to increase the whole operational efficiency of this kind of queueing system in real industrial applications. We further suggest that the management of companies should prepare repair processes for the system with possibility of happening breakdowns of service stations.

Future research will focus on statistical analysis of the real manufacturing systems and compare the results of the analysis with our theoretical results developed in this research. In addition, suggestions for the cases that the series configuration queueing system consisting of n service stations with different service rates are worth for future research. Transient analysis and the time to serve a customer follows general distributions would be considered further.

APPENDIX

The structure of the transition matrix Q and its sub-matrices for the system with two service stations

The transition matrix of the series configuration queueing system with two service stations can shown as

$$Q = \begin{bmatrix} B_{0,0} & A_0 & 0 & 0 & 0 & 0 & \cdots \\ A_2 & A_1 & A_0 & 0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ 0 & 0 & 0 & 0 & A_2 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

The following sub-matrices show the composition of the transition matrix corresponding to the quasi-birth-death process for the system with two service stations.

$$B_{0,0} = \begin{bmatrix} -\lambda & 0 & 0 \\ \mu_2 & -(\lambda+\mu_2) & 0 \\ 0 & \mu_2 & -(\lambda+\mu_2) \end{bmatrix},$$

$$A_0 = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & \mu_1 & 0 \\ 0 & 0 & \mu_1 \\ 0 & 0 & 0 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} -(\lambda+\mu_1) & 0 & 0 \\ \mu_2 & -(\lambda+\mu_1+\mu_2) & 0 \\ 0 & \mu_2 & -(\lambda+\mu_2) \end{bmatrix}$$

The structure of the transition matrix Q and its sub-matrices for the system with three service stations

$$Q = \begin{bmatrix} B_{0,0} & B_{0,1} & 0 & 0 & 0 & 0 & \cdots \\ B_{1,0} & A_1 & A_0 & 0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ 0 & 0 & 0 & 0 & A_2 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

The details of sub-matrices of the composition of the transition matrix corresponding to the quasi-birth-death process for the system with three service stations are given by

$$B_{0,0} = \begin{bmatrix} -\lambda & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -(\lambda+\mu_1) & 0 & 0 & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -(\lambda+\mu_2) & 0 & 0 & \lambda & 0 & 0 & 0 & 0 \\ \mu_3 & 0 & 0 & -(\lambda+\mu_3) & 0 & 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -(\lambda+\mu_1) & \mu_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -(\lambda+\mu_1+\mu_2) & \mu_2 & 0 & 0 & \mu_1 \\ 0 & \mu_3 & 0 & 0 & 0 & 0 & -(\lambda+\mu_1+\mu_3) & \mu_1 & 0 & 0 \\ 0 & 0 & \mu_3 & 0 & 0 & 0 & 0 & -(\lambda+\mu_2+\mu_3) & \mu_2 & 0 \\ 0 & 0 & 0 & \mu_3 & 0 & 0 & 0 & 0 & -(\lambda+\mu_3) & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_2 & 0 & -(\lambda+\mu_2) \end{bmatrix},$$

$$B_{1,0} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mu_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_3 \end{bmatrix},$$

$$B_{0,1}=\begin{bmatrix}0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\\lambda&0&0&0&0&0&0&0\\0&\lambda&0&0&0&0&0&0\\0&0&\lambda&0&0&0&0&0\\0&0&0&\lambda&0&0&0&0\\0&0&0&0&\lambda&0&0&0\\0&0&0&0&0&0&\lambda&0\end{bmatrix},$$

$$A_0=\begin{bmatrix}\lambda&0&0&0&0&0&0&0\\0&\lambda&0&0&0&0&0&0\\0&0&\lambda&0&0&0&0&0\\0&0&0&\lambda&0&0&0&0\\0&0&0&0&\lambda&0&0&0\\0&0&0&0&0&\lambda&0&0\\0&0&0&0&0&0&\lambda&0\\0&0&0&0&0&0&0&\lambda\end{bmatrix},$$

$$A_2=\begin{bmatrix}0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0\\\mu_3&0&0&0&0&0&0&0\\0&\mu_3&0&0&0&0&0&0\\0&0&\mu_3&0&0&0&0&0\\0&0&0&\mu_3&0&0&0&0\\0&0&0&0&0&0&0&0\\0&0&0&0&0&0&\mu_3&0\end{bmatrix},$$

$$A_1=\begin{bmatrix}-(\lambda+\mu_1)&\mu_1&0&0&0&0&0&0\\0&-(\lambda+\mu_1+\mu_2)&\mu_2&0&0&0&\mu_1&0\\0&0&-(\lambda+\mu_1+\mu_3)&\mu_1&0&0&0&0\\0&0&0&-(\lambda+\mu_1+\mu_2+\mu_3)&\mu_2&0&0&\mu_1\\0&0&0&0&-(\lambda+\mu_1+\mu_3)&\mu_1&0&0\\0&0&0&0&0&-(\lambda+\mu_3)&0&0\\0&0&0&\mu_2&0&0&-(\lambda+\mu_2)&0\\0&0&0&0&0&\mu_2&0&-(\lambda+\mu_2+\mu_3)\end{bmatrix}$$

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers 25287026 and 15K17583.

REFERENCES

[1] Road Bureau, Ministry of Land, Infrastructure, Transport and Tourism, Japan . (2006). Methods for solving traffic jams [Online]. Available: http://www.mlit.go.jp/road/ir/ir-perform/h18/07.pdf
[2] G.C. Hunt, "Sequential arrays of waiting lines," Operations Research, vol. 4, no. 6, pp. 674–683, 1956.
[3] B. Avi-Itzhak and M. Yadin, "A sequence of two servers with no intermediate queue," Management Science, vol. 11, no. 5, pp. 553-564, 1965.
[4] B. Avi-Itzhak and M. Yadin, "A sequence of service stations with arbitrary input and regular service times," Management Science, vol. 11, no. 5, pp. 565-571, 1965.
[5] T. Altiok, "Approximate analysis of exponential tandem queues with blocking," European Journal of Operational Research, vol. 11, no. 4, pp. 390-398, 1982.
[6] C. Langaris and B. Conolly, "On the waiting time of a two stage queueing system with blocking," Journal of Applied Probability, vol. 21, no. 3, pp. 628-638, 1984.
[7] H. T. Papadopoulos and M.E.J. O'Kelly, "Exact analysis of production lines with no intermediate buffers," European Journal of Operational Research, vol. 65, no. 1, pp. 118-137, 1993.
[8] B. Avi-Itzhak and S. Halfin, "Servers in tandem with communication and manufacturing blocking," Journal of Applied Probability, vol. 30, no. 2, pp. 429-437, 1993.
[9] I.F. Akyildiz and Horst von Brand, "Exact solutions for networks of queues with blocking-after-service," Theoretical Computer Science, vol. 125, no. 1, pp. 390-398, 1994.
[10] B. Avi-Itzhak and H. Levy, "A sequence of servers with arbitrary input and regular service times revisited," Management Science, vol. 41, no. 6, pp. 1039-1047, 1995.
[11] M.F. Neuts, Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. New York: Dover Publications, 1995, ch. 1.
[12] A. Gomez-Corral, "A matrix-geometric approximation for tandem queues with blocking and repeated attempts," Operations Research Letters, vol. 30, no. 6, pp. 360-374, 2002.
[13] A. Gomez-Corral, "On a tandem G-network with blocking," Advances in Applied Probability, vol. 34, no. 3, pp. 626-661, 2002.
[14] A. Gomez-Corral, "A tandem queue with blocking and Markovian arrival process," Queueing Systems, vol. 41, no. 4, pp. 343-370, 2002.
[15] A. Gomez-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.
[16] A. Gomez-Corral and M.E. Martos, "Marked Markovian arrivals in a tandem G-network with blocking," Methodology and Computing in Applied Probability, vol. 11, no. 4, pp. 621-649, 2009.
[17] R. Bierbooms, I.J.B.F. Adan, and M. van Vuuren, "Performance analysis of exponential production lines with fluid flow and finite buffers," IIE Transactions, vol. 44, no. 12, pp. 1132-1144, 2012.
[18] 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, vol. 209, no. 1, pp. 67-84, 2013.
[19] W. Zhou, Z. Lian, W. Xu, and W. Huang, "A two-stage queueing network with MAP inputs and buffer sharing," Applied Mathematical Modelling, vol. 37, no. 6, pp. 3736-3747, 2013.
[20] M. Hiller, "Designing unpaced production lines to optimize throughput and work-in-process inventory," IIE Transactions, vol. 45, no. 5, pp. 516-527, 2013.
[21] Y. Sakuma and A. Inoie, "An approximation analysis for an assembly-like queueing system with time-constraint items," Applied Mathematical Modelling vol. 38, no. 24, pp. 5870-5882, 2014.
[22] Y.W. Shin and D.H. Moon, "Approximation of throughput in tandem queues with multiple servers and blocking," Applied Mathematical Modelling, vol. 38, no. 24, pp. 6122-6132, 2014.
[23] S. Hudson, T. McNamara, and S. Shaaban, "Unbalanced lines: where are we now," International Journal of Production Research, vol. 53, pp. 1895-1911, 2015.
[24] G. Bolch and S. Greiner, Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications. New Jersey: Wiley-Interscience, 2nd edition, 2006, ch. 3.
[25] S. Sani and O.A. Daman, "The M/G/2 queue with heterogeneous servers under a controlled service discipline: stationary performance analysis," IAENG International Journal of Applied Mathematics, vol. 45, no. 1, pp. 31-40, 2015.
[26] S. Ramasamy, O.A. Daman, and S. Sani, "Discrete-time Geo/G/2 queue under a serial and parallel queue disciplines," IAENG International Journal of Applied Mathematics, vol. 45, no. 4, pp. 354-363, 2015.
[27] Y.L. Tsai, D. Yanagisawa, and K. Nishinari "Performance analysis of series configuration queueing system with four service stations," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2016, IMECS 2016, 16-18 March, 2016, Hong Kong, pp. 931–935.
[28] Y.L. Tsai, D. Yanagisawa, and K. Nishinari, "Performance analysis of open queueing networks subject to breakdowns and repairs," Engineering Letters, vol. 24, no. 2, pp. 207-214, 2016.
[29] H.G. Perros, "A bibliography of papers on queueing networks with finite capacity queues," Performance Evaluation, vol. 10, no. 3, pp. 255-260, 1989.
[30] H.G. Perros, Open Queueing Networks with Blocking. North-Holland: Amsterdam, 1989.
[31] R.O. Onvural, "Survey of closed queueing networks with blocking," ACM Computing Surveys, vol. 22, no. 2, pp. 83-121, 1990.
[32] S. Balsamo, "Properties and analysis of queueing network models with finite capacities," Performance Evaluation of Computer and Communication Systems, vol. 729, pp. 21-52, 2005.
[33] N.G. Hall and C. Sriskandarajah, "A survey of machine scheduling problems with blocking and no-wait in process," Operations Research, vol. 44, no.3, pp. 510-525, 1996.

[34] H.G. Perros, Queueing Networks with Blocking: Exact and Approximate Solutions. New York: Oxford University Press, 1994.
[35] S. Balsamo, V. de Nitto Persone, and Onvural, Analysis of Queueing Networks with Blocking. Boston: Kluwer Academic Publishers, 2001.
[36] G. Latouche and M.F. Neuts, “Efficient algorithmic solutions to exponential tandem queues with blocking,” SIAM Journal on Algebraic and Discrete Methods, vol. 1, no. 1, pp. 93-106, 1980.
[37] H.G. Perros and T. Altiok, Queueing Networks with Blocking. North-Holland: Amsterdam, 1989.
[38] F.S. Hillier and R.W. Boling, “Finite queues in series with exponential or Erlang service times – a numerical approach,” Operations Research, vol. 15, no. 2, pp. 286-303, 1967.
[39] H.G. Perros and T. Altiok, “Approximate analysis of open networks of queues with blocking: tandem configurations,” IEEE Transactions on Software Engineering, vol. 33, no. 3, pp. 450-461, 1986.