Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

DEPENDENCE AMONG TANDEM QUEUES AND THE SECOND MOMENT RESULT ON THE THEORY OF CONSTRAINTS

Kan Wu

Cluster of MIE
Nanyang Technology University
102 Nanyang Crescent
Singapore 637820, SINGAPORE

Ning Zhao

Faculty of Science
Kunming University of Science and Technology
68 Wenchang Road
Kunming 650093, CHINA

ABSTRACT

Due to the analytical intractability of general tandem queues, we develop innovative methods to quantify the dependence among stations through simulation. Dependence is defined by the contribution queue time at each station, and contribution factors are developed based on the insight from Friedman's reduction method and Jackson networks. In a tandem queue, the dependence among stations can be either diffusion or blocking, and their impact depends on the positions relative to the bottlenecks. Based on these results, we show that improving the performance of the system bottleneck may not be the most effective place to reduce system cycle time. Rather than making independence assumptions, the proposed method points out a promising direction and sheds light on the insights of the dependence in practical systems.

1 INTRODUCTION

To make an organization more profitable, production systems are often required by the management to have higher throughput rate under limited resource especially during peak seasons. To achieve this goal, Goldratt and Cox (1992) proposed the Theory of Constraints (TOC) based on the concept of bottlenecks, where a bottleneck is defined as the workstation whose required throughput rate is higher than its capacity. Through TOC, they explained how to achieve higher system throughput rate by relieving the bottleneck as well as how to reduce inventory by synchronizing production lines with the bottleneck. Rahman (1998) gave a comprehensive review on the Theory of Constraints

Stochastic effects are inherent in production systems: a machine faces different types of preventive maintenances, product changeovers or breakdowns. They can be either time-based, or run-based and preemptive or non-preemptive (Wu 2014a; Wu et al. 2011). A flexible machine can process different products with different service times under complicated dispatching policies. The transportation time may not be a constant between workstations and service time may not be the same as process time (Wu and Hui 2008; Wu et al. 2007). While service time variability can be small (Bitran and Tirupati 1988; Inman 1999), production environment is stochastic by nature.

In a stochastic system, the price of higher throughput rate is longer queue time. When the throughput rate approaches capacity, the queue time goes to infinity. Since no customer would accept infinite cycle time, a bottleneck defined by throughput rate cannot occur in a stochastic production line. On the other hand, the bottleneck in manufacturing is typically defined as the workstation with the highest level of utilization (see e.g. Lozinski and Classey, 1988 and Hopp and Spearman, 1995). However, due to the dependence among workstations, the station with the highest utilization may not have the most impact on system cycle time. To overcome the shortcomings, Wu (2005) extended the definition of bottlenecks from throughput bottlenecks (TPBN) to cycle time bottlenecks (CTBN), where a cycle time bottleneck is the workstation which prevents a production system from achieving its mean cycle time target. Since system

cycle time is contributed by all workstations, all workstations are cycle time bottlenecks with different levels of contribution. With the same mean cycle time target, reducing the mean queue time of any workstation would allow queue time increases of the others, and potentially lead to a higher throughput rate of the system under the same system capacity. By defining bottlenecks from the view point of cycle time, the concept of TOC has been extended from a deterministic system to a stochastic one. Although all workstations can be cycle time bottlenecks, there are still major and minor ones, where a major one has a higher impact on system mean cycle time. The question becomes which workstation is the major cycle time bottleneck and we should improve first?

For a workstation with independent and identically distributed (iid) interarrival time and service time, its mean queue time can be approximated by Kingman's *G*/*G*/1 heavy traffic approximation (Kingman 1965):

$$QT(G/G/1) \cong \left(\frac{c_a^2 + c_s^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)\frac{1}{\mu}, \tag{1}$$

where ρ is utilization ($= \lambda/\mu$), λ is arrival rate, μ is service rate, c_a^2 is the squared coefficient of variation (SCV) of arrival intervals, c_s^2 is the SCV of service time, and QT is mean queue time. Cycle time is the sum of queue time and service time. In a queueing network, if all stations work independently, Kingman's approximation would give good evaluation of system performance. However, in practice, congestion at a workstation often implies later congestion at its downstream stations. Machine states are dependent and the internal arrival process is not renewal in general (Whitt 1995). In this situation, the performance of a workstation will have impact on its downstream workstations. Simply improving the workstation with the highest utilization may not be the optimal choice. In terms of cycle time reduction, we call a workstation the first moment CTBN, if it is the most effective workstation to improve system cycle time when its service time is reduced, and we call a workstation the second moment CTBN if it is the most effective workstation to improve system cycle time when its variability (or variations) is reduced.

Although the existence of dependence among stations is well recognized, due to the non-renewal departure processes (Bitran and Dasu 1992), the exact analysis of dependence in general queueing systems is analytically intractable. The current approaches to evaluate the performance of queueing networks are mainly based on independence assumptions directly (e.g. the stochastic independence assumption (Kleinrock 1976)) or indirectly (e.g. the functional central limit theorem (Harrison and Nguyen 1990)). Due to the independence assumptions, product-form and Brownian networks are not capable to fully capture the dependence among stations.

To have better understanding of practical queueing systems, we study the dependence of mean queue times among stations in general tandem queues through simulation. Dependence is defined by the contribution of a station in a tandem queue, and the contribution of a station is defined based on the insight from Friedman's reduction method (Friedman 1965) and Jackson networks (Jackson 1957). Two types of dependences are identified: blocking and diffusion effects. Their impact on system queue time depends on their positions relative to the bottlenecks. We start our investigation from a simple problem with the following assumptions: workstations are arranged in series without reentry, each workstation is a single server with infinite buffers, dispatching policy is first-come-first-server (FCFS), and the service times of each workstation and the external interarrival times are mutually independent and generally distributed.

This paper is organized as follows. Section 2 reviews the property of intrinsic ratios and defines the contribution queue times. Section 3 explains the dependence among single server queues in series. Section 4 introduces the second moment results on the theory of constraints, and conclusion is given in Section 5.

2 INTRINSIC RATIO AND CONTRIBUTION QUEUE TIME

In this study we investigate the dependence of the mean queue times of a general tandem queue with N single server stations as shown in Figure 1. The external interarrival times and service times are mutually independent and generally distributed. Jobs arrive at the first station independently with arrival rate λ and squared coefficient of variation (SCV) c_a^2. There are infinite buffers at each station and the service discipline is first-come first-served (FCFS). Denote the service time at station i by S_i, and SCV of S_i by $c_{S_i}^2, i = 1, \cdots, N$. Let service rate at station i be μ_i and $\rho_i = \lambda/\mu_i < 1$.

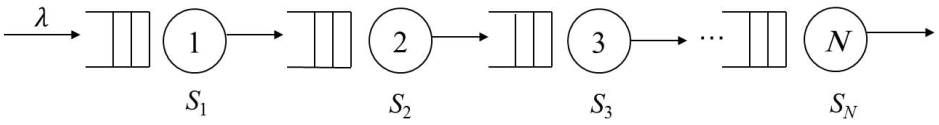

Figure 1: Tandem queues with N single server stations.

Wu and McGinnis (2013) studied tandem queues in Figure 1 and introduced the concept of intrinsic ratio. They also proposed an approximate model for the system queue time of a general queueing network through intrinsic ratios (Wu and McGinnis 2012; Wu and Zhao 2015a). Here we give a brief review of the intrinsic ratio and system queue time approximation. It constitutes the fundamentals of the analysis in Section 3. To compute the intrinsic ratio, bottlenecks of a tandem queue have to be determined first as follows.

Procedure 1: Identification of bottlenecks

1. Identify the index of the main system bottleneck server (BN_1), where $\mu_{BN_1} = \min \mu_i$, for $i = 1$ to N.

   Let $k = 1$.

   - If more than one server has the minimum service rate, $BN_1 = \min i$, where $\mu_i = \mu_{BN_1}$.
2. Identify the index of the next bottleneck server (BN_{k+1}) in front of the previous one (BN_k), where $\mu_{BN_{k+1}} = \min \mu_i$, for $i = 1$ to $BN_k - 1$.

   - If more than one server has the minimum service rate, $BN_{k+1} = \min i$, where $\mu_i = \mu_{BN_{k+1}}$.
3. If $BN_{k+1} = 1$, then go to step 4. Otherwise, let $k = k + 1$ and go to step 2.
4. Stop.

Procedure 1 identifies the main system bottleneck first, and then identifies the next bottleneck within a subsystem, where a subsystem is composed of the servers from the first server to the newest identified bottleneck (not included). At first when no bottleneck has been identified, the subsystem is the entire system and BN_1 is the system bottleneck. The subsystem then gradually becomes smaller until the subsystem is solely composed of the first station of the tandem queue.

To compute intrinsic ratios, Wu and McGinnis (2013) introduced ASIA and fully coupled systems. In an ASIA system, all servers see the initial arrivals (ASIA) directly. Therefore an ASIA system of station i is a $G/G/1$ queue with the initial arrival process and service time S_i ($1 \le i \le N$) as shown in Figure 2.

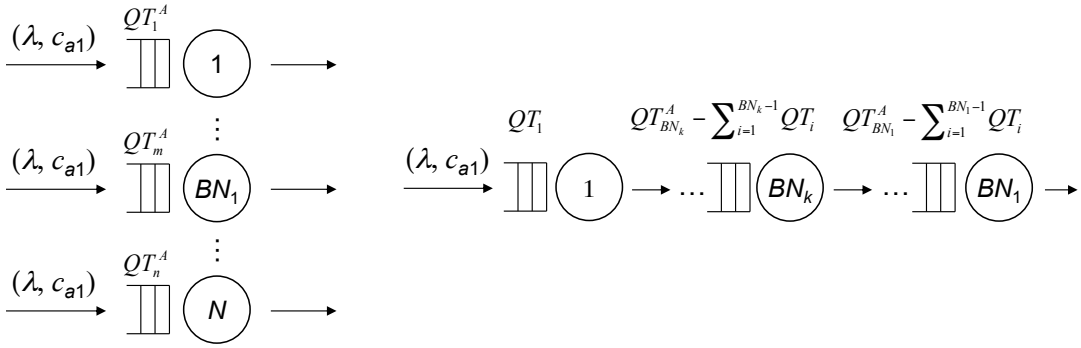

Figure 2: Mean queue times in the ASIA (left) and fully coupled (right) systems.

Definition 1 In a*n ASIA system, the mean queue time of the i-th server* QT_i^A *is the mean queue time of the server when it sees the initial arrivals directly.*

Based on Kingman's *G*/*G*/1 heavy traffic approximation (Kingman 1965), the mean queue time (QT_i^A) of station i in an ASIA system can be approximated by

$$QT_i^A \approx \frac{c_{a1}^2 + c_{S_i}^2}{2} \frac{\rho_i}{1-\rho_i} \frac{1}{\mu_i}, \qquad 1 \le i \le N\,. \tag{2}$$

An ASIA system gives the exact result when all service times are exponentially distributed and the external arrival process is Poisson. In contrast with an ASIA system, motivated by Friedman's reduction method, a fully coupled system is defined as follows.

Definition 2 *In a fully coupled system, all non-bottleneck servers have zero queue times and the mean queue time of the k-th bottleneck* $QT_{BN_k}^C = QT_{BN_k}^A - \sum_{i=1}^{BN_k-1} QT_i$*, where* $QT_{BN_k}^A$ *is the mean queue time at the k-th bottleneck when it sees the initial arrivals directly and* QT_i *is the mean queue time of the i-th server in the original tandem queue.*

In a fully coupled system, the system mean queue time is purely dominated by a single station (i.e., the bottleneck). It is true for a tandem queue when all service times are constant (Friedman 1965). The mean queue time (QT_i^C) of station i in a fully coupled system is

$$QT_i^C = \begin{cases} QT_i, & \text{if } i = 1, \\ QT_i^A - \sum_{k=1}^{i-1} QT_k\,, & \text{if } i \ge 2 \text{ and station } i \text{ is a bottleneck}, \\ 0, & \text{if } i \ge 2 \text{ and station } i \text{ is not a bottleneck}, \end{cases} \tag{3}$$

where QT_i is the mean queue time at station i in the original tandem queue. Note that by Eq. (3), the summation of all mean queue times in a fully coupled system is $QT_{BN_1}^A$. Since QT_i^C is motivated by a tandem queue with service time SCV = 0 and QT_i^A is motivated by a tandem queue with service time SCV = 1, it is important to examine how the queue time of a tandem queue with different service time SCVs performs relative to these two bounds. The intrinsic ratio of station i is defined as

$$r_i = \frac{QT_i - QT_i^C}{QT_i^A - QT_i^C}\,, \quad 2 \le i \le N. \tag{4}$$

In a Jackson network (Jackson 1957), since all servers see Poisson arrivals and behave independently in steady state, $QT_i = QT_i^A$ and the intrinsic ratio is one at all stations based on Eq. (4). In a tandem queue, when the service times of all stations are constant (or a Friedman's tandem queue), system mean queue time is solely determined by the main system bottleneck based on the reduction method (Friedman 1965). Namely, the system mean queue time of the tandem queue is the same as the mean queue time of the main system bottleneck when it sees the initial external arrival process directly, i.e., $\sum_{i=1}^{N} QT_i =$

$QT^A_{BN_1}$, where $QT^A_{BN_1}$ is the mean queue time of the main system bottleneck in the ASIA system. Since the service times of all stations are constant, the equality indeed holds in all sub-systems (i.e., replacing BN_1 by BN_k, and N by the total number of servers in the sub-system.), and the non-bottleneck stations after the bottleneck have zero queue times in each sub-system. The equality is equivalent to $\sum_{i=1}^{BN_k} QT_i = QT^A_{BN_k}$. In this situation, $QT_i = QT_i^C$ by Eq. (3) and the intrinsic ratio is zero at all stations.

Because Eq. (2) reduces to an M/G/1 queue (and gives exact results) when the arrival process is Poisson, the computation of intrinsic ratio is accurate with Poisson arrivals. With other general arrival processes, in order to get the correct intrinsic ratio, the mean queue time has to be obtained from simulation. Wu and McGinnis (2013) observed that the intrinsic ratio behaves approximately linear across traffic intensities. Based on this nice property, Wu and McGinnis (2012) derived Eq. (5) to approximate system mean queue time (QT_f) of *N* servers in series.

$$QT_f = \sum_{i=1}^{N} QT_i = \sum_{i=1}^{N} f_{N,i} * QT_i^A, \tag{5}$$

where $f_{N,i}$ is called contribution factor (CF) and can be determined by Procedure 2.

Procedure 2: Determining the parameters $f_{N,i}$

1. Let $k = N$, $f_{N,i} = 1$ for $i = 1$ to *N*.
2. If server k is marked as a bottleneck, $f_{N,i} = r_k * f_{N,i}$ for $i = 1$ to $k - 1$.

Otherwise, $f_{N,k} = r_k * f_{N,k}$. Stop if $k = 2$.

3. Let $k = k - 1$, and go to step 2.

In a Jackson network (Jackson 1957), since the intrinsic ratio is one for all stations, contribution factors equal one at all stations as well. In a Friedman's tandem queue, since all intrinsic ratios are zero, the contribution factor is zero at the non-bottleneck stations, and is one at the bottleneck based on Procedure 2.

According to Procedure 2, the value of CF is always one at the main system bottleneck, but can be other non-negative real numbers at the non-bottlenecks. An important observation is that the ASIA system plays a critical role in determining the contribution of a server to the entire system. In a Jackson network, all servers see the initial arrivals, which is a Poisson process (Burke 1956). Hence, each station behaves as if it is a standalone station, which is also its counterpart in the ASIA system. In this situation, the ASIA system mean queue time of each station is the same as (or "contributes" to) the actual mean queue time of the original system.

In a tandem queue with constant service times at all stations, the system mean queue time is solely determined by the main system bottleneck according to the reduction method (Friedman 1965). The main system bottleneck station behaves as if it is a standalone station, which sees the initial arrivals directly. In this situation, the ASIA system queue time of the bottleneck determines (or "contributes" to) the system queue time, but all the non-bottlenecks have no impact or contribution to the overall system queue time, despite that jobs do have *appeared queue time* at those non-bottleneck stations, where appeared queue time refers to the actual time a customer spends waiting at a particular station.

Let $C_{N,i} = f_{N,i} * QT_i^A$. Relative to the mean queue time in its ASIA system, $C_{N,i}$ is the contribution from station i to the system mean queue time. Based on the above analysis, there can be two types of queue times associated with each station in a tandem queue: QT_i and $C_{N,i}$. QT_i is the actual or appeared queue time at station i, which measures its performance in appearance. $C_{N,i}$ is the contribution queue time at station i, which measures its authentic contribution to the system. Based on Eq. (5), the total appeared queue time is the same as the total contribution queue time. When the service times of all stations in a tandem queue are constant, although jobs may have queue times at the non-bottleneck stations (i.e.,

positive appeared queue time), those non-bottleneck stations indeed have zero contribution to the system queue time, since the system queue time is solely determined by the main system bottleneck and those non-bottleneck stations play no role in determining system queue time (Friedman 1965).

The contribution queue time coincides with the appeared queue time in a Jackson network, but they should be different in general. In contrast to the appeared queue time, although the contribution queue time is more abstract, it represents the true contribution of a station to the system. It plays the key role to analyze the dependence among the stations of a tandem queue.

3 DEPENDENCE AMONG SINGLE SERVER QUEUES IN SERIES

In this section, we study the dependence from the viewpoint of ASIA systems, and show how to describe the dependence among stations through the concept of contribution queue time. In the following analysis, we use the single-point historical data approach (Wu and McGinnis 2013), which assumes the historical queue time of each workstation is known at a specific traffic intensity. In our case, we get the historical data from simulations. r_i can be obtained through the simulation queue time and Eqs. (2) to (4).

The FCFS policy only considers local information and is a decentralized control policy. Under FCFS, in a tandem queue with infinite buffer capacity, the states of downstream stations have no impact on the dispatching decisions (and hence the arrival processes) of the upstream servers. When the service times of servers and the external interarrival times are mutually independent, the above observation implies the states of the downstream stations have no impact on the queue times of the upstream stations. On the other hand, since the arrival process of a downstream station is the departure process of its upstream stations, the queue time of a downstream station is dependent on the states of its upstream servers.

Since the actual (or appeared) queue time of an upstream station is not impacted by its downstream workstations (but the contribution queue time can be), dependence among stations is analyzed based on the changes of the contribution queue times from the first station to the last one by sequentially adding the stations back to the system (i.e., increasing the value of N in Eq. (5)). When $N = k - 1$ ($k \geq 2$), the contribution factor of station i in the subsystem which is composed of the first $k-1$ servers is $f_{k-1,i}, i = 1, \cdots, k-1$. The contribution queue time from station i is $C_{k-1,i} = f_{k-1,i} * QT_i^A$. Similarly, when $N = k$, the contribution factor of station i in the subsystem composed of the first k servers is $f_{k,i}, i = 1, \cdots, k$. The contribution queue time of station i is $C_{k,i} = f_{k,i} * QT_i^A$. Let $X(i,k) = C_{k,i} - C_{k-1,i}$. If station k is a bottleneck in Procedures 1 and it is added to a subsystem which is composed of the first $k-1$ stations of the tandem queue, $X(i,k)$ measures the change of the contribution queue time of station i due to the existence of station k.

- If $X(i,k) > 0$, the actual mean queue time at station k is increased by $X(i,k)$ due to station i. Namely, the diffused queue time at station k from station i is $X(i,k)$. We call this phenomenon diffusion on bottleneck (DoB).
- If $X(i,k) < 0$, the actual mean queue time at station k is decreased by $-X(i,k)$ due to station i. Namely, the blocked queue time at station k from station i is $-X(i,k)$. We call this phenomenon blocking on bottleneck (BoB).

If station k is a bottleneck, the actual (or appeared) queue time at station k is $QT_k = QT_k^A + \sum_{i=1}^{i=k-1} X(i,k)$. On the other hand, if station k is a non-bottleneck in Procedures 1, based on Procedure 2, the contribution queue time of station i doesn't change when adding the kth station to the subsystem which is composed of the first $k-1$ stations of the tandem queue. The contribution queue time of station k is $C_{k,k} = f_{k,k} * QT_k^A$. Let $X(1\colon k-1, k) = C_{k,k} - QT_k^A$.

- If $X(1:k-1,k) > 0$, the actual mean queue time at station k is increased by $X(1:k-1,k)$ due to station 1 to station $k-1$. Namely, the diffused queue time at station k from stations 1 to station $k-1$ is $X(1:k-1,k)$. We call this phenomenon diffusion on non-bottleneck (DoN).
- If $X(1:k-1,k) < 0$, the actual mean queue time at station k is decreased by $-X(1:k-1,k)$ due to station 1 to station $k-1$. Namely, the blocked queue time at station k from station 1 to station $k-1$ is $-X(1:k-1,k)$. We call this phenomenon blocking on non-bottleneck (BoN).

Hence, if station k is a non-bottleneck, the actual (or appeared) queue time at station k is $QT_k = QT_k^A + X(1:k-1,k)$. In the following, we present a queueing systems to explain BoB. The system is composed of five single server queues in series. For each simulation experiment, thirty replications are conducted at the specific arrival rate. Each replication consists of 200,000 jobs after discarding the first 400,000 jobs for warm-up. The sample size is sufficiently large so that the half width of 95% confidence intervals of the mean simulation queue time is less than 2%.

3.1 Blocking on Bottlenecks (BoB)

To illustrate the blocking effects on bottlenecks, a case where the service time SCVs are smaller than the initial interarrival time SCV is investigated. The system is composed of five single server queues in series. The mean service time from the first server to the last is 20, 23, 25, 28 and 30 min. The service time SCVs are 0.25 for all servers. The arrival process is Poisson with mean 33 1/3 min.

Based on the simulation results, the actual queueing times are $QT_1 = 18.75$, $QT_2 = 22.88$, $QT_3 = 30.37$, $QT_4 = 63.47$ and $QT_5 = 116.73$. The ASIA system queue times of the five stations are $QT_1^A = 18.75$, $QT_2^A = 32.00$, $QT_3^A = 46.88$, $QT_4^A = 91.88$ and $QT_5^A = 168.75$. By Eq. (4) the intrinsic ratios are r_2=0.5136, r_3=0.6034, r_4=0.6054 and r_5=0.6160.

The blocking effect is analyzed by assuming $\mathrm{N} = 1$ in Eq. (5) and then gradually increasing N one by one. For example, when $\mathrm{N} = 2$, $\mathrm{C}_{2,1} = 9.63$ and $\mathrm{X}(1,2) = -9.12$. Hence, the contribution queue time at station 1 is 9.63, even if its actual queue time is 18.75. The difference of 9.12 (i.e., $-\mathrm{X}(1,2)$) is the portion of station 2's ASIA system queue time blocked by station 1. Hence, the actual mean queue time at station 2 is $\mathrm{QT}_2 = \mathrm{QT}_2^{\mathrm{A}} + \mathrm{X}(1,2) = 22.88$. Information of subsystems is listed in Table 1. The blocking effect in this tandem queue is demonstrated in Figure 3.

Table 1: Analysis for the BoB case.

		k=1	k=2	k=3	k=4	k=5
N=1	$f_{1,k}$	1				
	$C_{1,k}$	18.75				
N=2	$f_{2,k}$	0.5136	1			
	$C_{2,k}$	9.63	32.00			
	$X(k,2)$	-9.12				
N=3	$f_{3,k}$	0.3099	0.6034	1		
	$C_{3,k}$	5.81	19.31	46.88		
	$X(k,3)$	-3.82	-12.69			
N=4	$f_{4,k}$	0.1876	0.3653	0.6054	1	
	$C_{4,k}$	3.52	11.69	28.38	91.88	
	$X(k,4)$	-2.29	-7.62	-18.50		
N=5	$f_{5,k}$	0.1156	0.2250	0.3729	0.6160	1
	$C_{5,k}$	2.17	7.20	17.48	56.59	168.75
	$X(k,5)$	-1.35	-4.49	-10.90	-35.28	

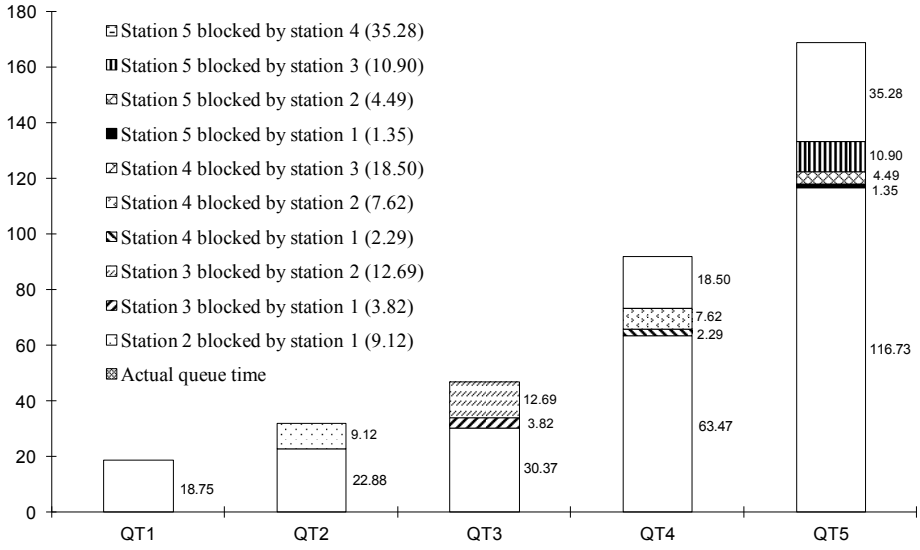

Figure 3: BoB analysis for the entire tandem queue.

In this tandem queue, the service time SCVs of all stations are 0.25, which are smaller than the SCV of the external arrival process. Hence, each station behaves like a breakwater, which lessens the departure variations from its upstream station. As a result, the actual queue time of the downstream station is shorter than its ASIA system queue time. The different between the actual queue time and ASIA system queue time is "blocked" by the upstream stations, and the queue time contribution from an upstream station to the system becomes less due to its blocking effect. Note that the contribution queue time is always the same as the ASIA system queue time at the main system bottleneck according to Procedure 2. Hence, except for the last station (which is the main system bottleneck), the contribution queue times of all other stations are shorter than the actual (or appeared) queue times. That is, their contributions are smaller than the appearance in this tandem queue system due to the blocking effect.

In this tandem queue, the service time SCVs of all stations are 0.25, which are smaller than the SCV of the external arrival process. Hence, each station behaves like a breakwater, which lessens the departure variations from its upstream station. As a result, the actual queue time of the downstream station is shorter than its ASIA system queue time. The different between the actual queue time and ASIA system queue time is "blocked" by the upstream stations, and the queue time contribution from an upstream station to the system becomes less due to its blocking effect. Note that the contribution queue time is always the same as the ASIA system queue time at the main system bottleneck according to Procedure 2. Hence, except for the last station (which is the main system bottleneck), the contribution queue times of all other stations are shorter than the actual (or appeared) queue times. That is, their contributions are smaller than the appearance in this tandem queue system due to the blocking effect.

As shown in Figure 3, the blocked queue time caused by the first station becomes less on the downstream stations when they are farther away from the first station in this examined case, i.e., $\mathrm{X}(1,2) = -9.12, \mathrm{X}(1,3) = -3.82, \mathrm{X}(1,4) = -2.29, \mathrm{X}(1,5) = -1.35$.

4 SECOND MOMENT RESULTS ON THE THEORY OF CONSTRAINTS

In practical manufacturing systems, cycle time reduction is often achieved through improving job scheduling or reducing the mean or variance of service times. According to Delp et al. (2006), reducing

service time variability is a cost-effective way to achieve shorter cycle time. Assuming the cost of reducing service time variability (in percentage) is the same for all stations, an important question to ask is how to identify the second moment CTBN, which is the most effective station to improve system cycle time if its variability is reduced.

The analysis of blocking and diffusion effects discussed in Section 3 presents a way to quantify the dependence among single server queues in series. When the variability of station i is reduced, the mean queue times at station i and its downstream stations are shortened. Denote the mean queue time of the original system by $QT_f(0,0)$. If the service time SCV of station i is reduced by p (in percentage), the new system mean queue time is denoted by $QT_f(i,p)$, $1 \le i \le N, 0 \le p \le 1$. If the mean service times of all stations are unchanged, the second moment CTBN is the station resulting in the shortest system mean queue time, i.e., $min_{1\le i\le N}\{QT_f(i,p)\}$.

Rather than improving the throughput bottleneck (with the highest utilization), we will show that it can be more effective to start with some other stations when one would like to reduce the system mean cycle time. An example is presented in the following. All the cases are examined at 10 different utilizations (ρ ranges from 0.1 to 0.95). Thirty replications are conducted at each utilization. Each replication consists of 200,000 jobs after discarding the first 400,000 jobs for warm-up. The sample size is sufficiently large so that the half width of 95% confidence intervals of the mean simulation queue time is less than 2%.

Example: A tandem queue consists of five single-server stations in series. The SCV of the external arrival interval at the first station is 0.3. The mean service times from the first station to the last are 28, 26, 24, 22 and 30 min. The service time SCVs from the 1st station to the last is SCV(0) = [0.8,0.8,0.8,0.8,0.8]. Let system utilization $\rho = \lambda / \mu_5$ and $p = 50\%$. The second moment CTBN is identified by reducing the service time SCV of each station. Let SCV(1) = [0.4,0.8,0.8,0.8,0.8], SCV(2) = [0.8,0.4,0.8,0.8,0.8], SCV(3) = [0.8,0.8,0.4,0.8,0.8], SCV(4) = [0.8,0.8,0.8,0.4,0.8] and SCV(5) = [0.8,0.8,0.8,0.8,0.4]. The system mean queue times of the six tandem queues with SCV(*i*), *i* = 0, 1, …, 5 are compared.

The system mean queue times from simulation are presented in Table 5. Among the six different configurations, the shortest queue time at each utilization is underlined. Based on the discussion in Section 3, because the SCV of the external arrival intervals is less than the SCVs of service times, diffusion effects exist in Example 1. The main system bottleneck is station 5 as μ_5 is the least among the five stations. Simulation results show that when $0.1 \le \rho \le 0.9$, the system mean queue time becomes the shortest if the variability of the first station (rather than the main system bottleneck) is reduced by 50%. The second moment CTBN is station 1.

Table 2: Identification of QTBN for the tandem queue.

ρ	$QT_f(0,0)$	$QT_f(1,0.5)$	$QT_f(2,0.5)$	$QT_f(3,0.5)$	$QT_f(4,0.5)$	$QT_f(5,0.5)$
0.1	1.63	1.33	1.38	1.43	1.48	1.42
0.2	7.90	6.62	6.88	7.07	7.24	7.05
0.3	18.36	15.57	16.17	16.64	17.04	16.59
0.4	33.23	28.48	29.54	30.35	30.97	30.20
0.5	54.12	46.39	48.27	49.52	50.58	49.15
0.6	84.08	72.13	75.15	77.27	78.79	76.28
0.7	129.37	111.49	116.70	119.78	122.08	116.80
0.8	208.70	179.53	188.95	194.80	197.70	186.96
0.9	388.79	336.96	357.86	368.96	374.53	340.62
0.95	652.69	573.47	610.73	625.14	635.92	548.87

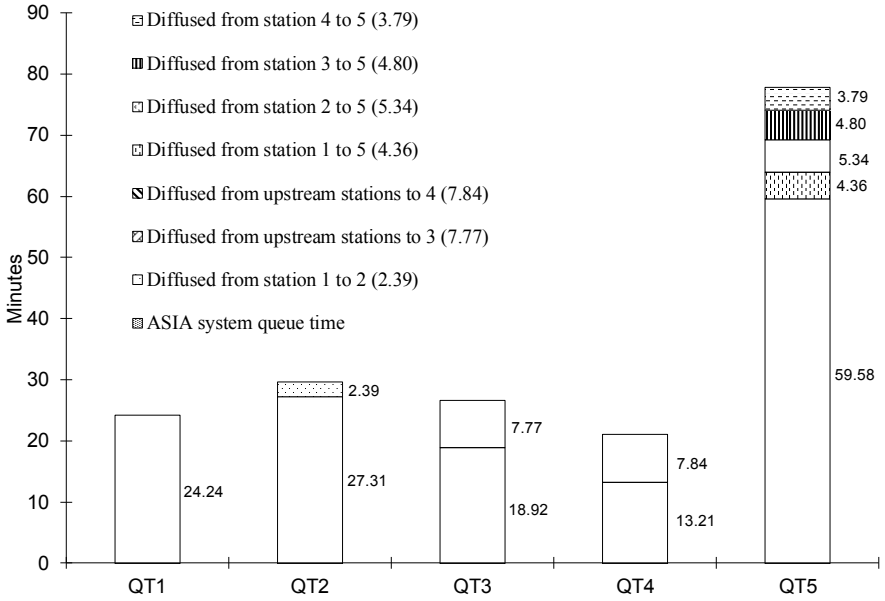

Figure 4: Queue time analysis for the tandem queue with SCV(1) and $\boldsymbol{\rho = 0.8}$.

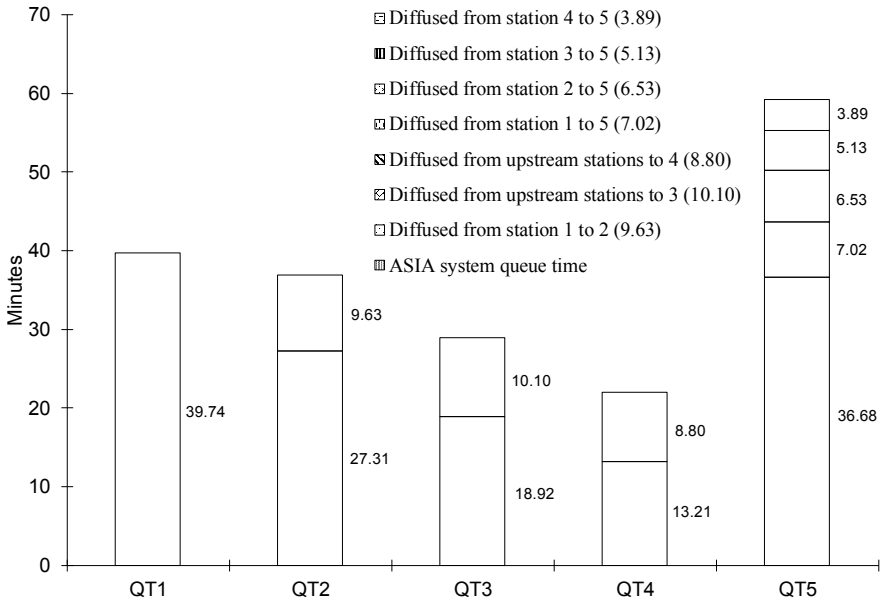

Figure 5: Queue time analysis for the tandem queue with SCV(5) and $\boldsymbol{\rho = 0.8}$.

To explore the reason, dependence among the tandem queues with SCV(1) and SCV(5) at $\rho = 0.8$ is analyzed. For the system with SCV(1), due to the diffusion effects, the reduction of variability at the first station not only reduces the queue time at station 1, but also reduces queue times at the downstream stations. However, in the system with SCV(5), the reduction of variability at the fifth station only reduces the queue time at station 5. As shown in Figures 10 and 11, the diffusion effect is less in the system with H(1), but is larger in the system with SCV(5). For example, the diffused queue time from station 1 to 2 is 2.39 under SCV(1), but it is 9.63 under SCV(5).

However, at $\rho = 0.95$, the system mean queue time is the shortest if the variability of station 5 is reduced by 50%. This is because the queue time at station 5 is much longer than the others, and system queue time is dominated by the main system bottleneck in heavy traffic. Shortening system queue time by reducing the service time variability of station 5 becomes more effective in this situation.

5 CONCLUSION

Although the existence of dependence among stations is well recognized, it is difficult to be modeled exactly in general and independence is commonly assumed directly or indirectly in queueing models. In order to improve the performance evaluation of queueing networks, accurately capturing the dependence into the mathematical model is of critical importance. This study can be viewed as a preliminary attempt towards this goal. Through contribution factors, the dependence among single server queues in series is analyzed. Due to the dependence among stations, productivity improvement should focus on not only the first moment CTBN, but also the second moment CTBN. While the first moment CTBN always dominates system queue time performance in heavy traffic, the second moment CTBN can play a more important role in the practical range of system utilizations.

In this study, we analyze the dependence among single stations in series with simple configuration. However, the station may have finite buffer capacity (Wu and Zhao 2015b) with process and arrival batches (Wu 2014b) or discretionary priority (Zhao et al. 2015). The process flow may have reentry. All these will make the analysis of dependence more complicated and the topic is left for future research.

Through contribution factors and simulation results, we find the dependence among workstations can be either blocking or diffusion. However, the conditions for blocking and diffusion effects have not been fully understood. For example, it is not clear that if the service time SCV less than the interarrival time SCV is a sufficient condition for the blocking effect. To understand the dependence better, rather than relying on simulation, computing intrinsic ratios analytically is necessary and left for future research.

ACKNOWLEDGMENTS

This research is supported in part by GSK-Singapore Partnership for Green and Sustainable Manufacturing under Grant M406884.

REFERENCES

Bitran, G., D. Tirupati. 1988. "Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference". *Management Scienc*e. 34(1) 75-100.

Bitran, G.R., S. Dasu. 1992. "A Review of Open Queueing Network Models of Manufacturing Systems". *Queueing Systems*. 12(1-2) 95-133.

Burke, P.J. 1956. "The Output of a Queuing System". *Operations Research*. 4(6) 699-704.

Delp, D., J. Si, J.W. Fowler. 2006. "The Development of the Complete X-Factor Contribution Measurement for Improving Cycle Time and Cycle Time Variability". *Semiconductor Manufacturing, IEEE Transactions on*. 19(3) 352-362.

Friedman, H.D. 1965. "Reduction Methods for Tandem Queuing Systems". *Operations Research*. 13(1) 121-131.

Goldratt, E.M., J. Cox. 1992. *The Goal: A Process of Ongoing Improvement*. North River Press.

Harrison, J.M., V. Nguyen. 1990. "The Qnet Method for Two-Moment Analysis of Open Queueing Networks". *Queueing Systems*. 6(1) 1-32.

Hopp, W.J., M.L. Spearman. 1995. *Factory Physics*. Irwin/McGraw-Hill Burr Ridge, IL.

Inman, R.R. 1999. "Empirical Evaluation of Exponential and Independence Assumptions in Queueing Models of Manufacturing Systems". *Production and Operations Management*. 8(4) 409-432.

Jackson, J.R. 1957. "Networks of Waiting Lines". *Operations Research*. 5(4) 518-521.

Kingman, J.F.C. 1965. "The Heavy Traffic Approximation in the Theory of Queues". in Proceedings of the *Symposium on Congestion Theory*, University of North Carolina at Chapel Hill, 137-159.

Kleinrock, L. 1976. *Queueing Systems: Volume 2: Computer Applications*. John Wiley & Sons.

Lozinski, C., C.R. Glassey. 1988. "Bottleneck Starvation Indicators for Shop Floor Control". *IEEE Transactions on Semiconductor Manufacturing*. 1(4) 147-153.

Rahman, S. 1998. "Theory of Constraints: A Review of the Philosophy and Its Applications". *International Journal of Operations & Production Management*. 18(4) 336-355.

Whitt, W. 1995. "Variability Functions for Parametric-Decomposition Approximations of Queueing Networks". *Management Science*. 41(10) 1704-1715.

Wu, K. 2005. "An Examination of Variability and Its Basic Properties for a Factory". *IEEE Transactions on Semiconductor Manufacturing*. 18(1) 214-221.

Wu, K. 2014a. "Classification of Queueing Models for a Workstation with Interruptions: A Review". *International Journal of Production Research*. 52(3) 902-917.

Wu, K. 2014b. "Taxonomy of Batch Queueing Models in Manufacturing Systems". *European Journal of Operational Research*. 237(1) 129-135.

Wu, K., K. Hui. 2008. "The Determination and Indetermination of Service Times in Manufacturing Systems". *IEEE Transactions on Semiconductor Manufacturing*. 21(1) 72-82.

Wu, K., L. McGinnis. 2012. "Performance Evaluation for General Queueing Networks in Manufacturing Systems: Characterizing the Trade-Off between Queue Time and Utilization". *European Journal of Operational Research*. 221(2) 328-339.

Wu, K., L. McGinnis. 2013. "Interpolation Approximations for Queues in Series". *IIE Transactions*. 45(3) 273-290.

Wu, K., L. McGinnis, B. Zwart. 2011. "Queueing Models for a Single Machine Subject to Multiple Types of Interruptions". *IIE Transactions*. 43(10) 753-759.

Wu, K., L.F. McGinnis, B. Zwart. 2007. "Compatibility of Queueing Theory, Manufacturing Systems and Semi Standards." in Proceedings of the *IEEE International Conference on Automation Science and Engineering*, 501-506.

Wu, K., N. Zhao. 2015a. "Dependence Among Single Stations in Series and its Applications in Productivity Improvement". *European Journal of Operational Research*. 247(1) 245-258.

Wu, K., N. Zhao. 2015b. "Analysis of Dual Tandem Queues with Finite Buffer Capacity and Nonoverlapping Service Times Subject to Breakdowns". *IIE Transactions*, DOI: 10.1080/0740817X.2015.1055389.

Zhao, N., Z. Lian, K. Wu. 2015. "Analysis of a MAP/PH/1 Queue with Discretionary Priority". *Asia-Pacific Journal of Operational Research*, DOI: 10.1142/S0217595915500426.

AUTHOR BIOGRAPHIES

KAN WU is an Assistant Professor at Nanyang Technological University. He received the Ph.D. degree in Industrial and Systems Engineering from Georgia Institute of Technology. He has 10 years' experience in the semiconductor industry, from consultants to managers. His PhD dissertation was awarded the 3rd place for the IIE Pritsker Doctoral Dissertation Award in 2010. His research interests are primarily in the areas of queueing theory, with applications in the performance evaluation of supply chains and manufacturing systems. His email is wukan@ntu.edu.sg.

NING ZHAO is an Assistant Professor in Faculty of Science at Kunming University of Science and Technology. She received the M.S. degree in Applied Mathematics from Shanghai Jiaotong University, and received a Ph.D in Business Information Systems from University of Macau. Her research interest is in queueing theory. Her email is zhaoning1102@gmail.com.