European Journal of Operational Research 247 (2015) 245–258

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Dependence among single stations in series and its applications in productivity improvement

Kan Wu [a,1], Ning Zhao [b,*]

[a] School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore
[b] Faculty of Science, Kunming University of Science and Technology, Kunming, China

ARTICLE INFO

Article history:
Received 18 September 2014
Accepted 9 May 2015
Available online 1 June 2015

Keywords:
Productivity
Simulation
Theory of constraint
Tandem queue

ABSTRACT

Theory of constraints has been commonly used in production systems to improve productivity. Since the improvement on an upstream workstation may have impact on its downstream servers, finding the true bottleneck is not trivial in a stochastic production line. Due to the analytical intractability of general tandem queues, we develop 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, McGinnis, & Zwart, 2011). A flexible machine can process different products with different service times under complicated dispatching policies. Jobs may encounter process and transfer batches (Wu, 2014b), and the transportation time may not be a constant between workstations. While service time variability can be small (Bitran & 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; 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

* Corresponding author. Tel.: +86 13888058515; fax: +86 871 65916345.
E-mail addresses: kan626@gmail.com (K. Wu), zhaoning@kmust.edu.cn (N. Zhao).
[1] Tel.: +65-6790-5584; fax: +65-6792-4062.

http://dx.doi.org/10.1016/j.ejor.2015.05.028
0377-2217/

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):

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

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.

Dependence in queueing systems has been widely studied since 1960s. Dependence among service times in tandem queues has been studied by Mitchell, Paulson, and Beswick (1977), Pinedo and Wolff (1982), Sandmann (2012), Weber and Weiss (1994) and Wolff (1982). Dependence between interarrivals and service time in queueing systems has been studied by Adan and Kulkarni (2003), Bhat (1969), Borst and Boxma (1993) and Boxma and Perry (2001). Prior literatures mainly focused on the dependence among successive service times, dependence between interarrival times and service times, as well as dependence among successive interarrival times (Fendick, Saksena, & Whitt, 1989). Only few of them studied the dependence among queue times. Reich (1963) proved that in single-server tandem queues with exponential service time and Poisson arrival process, the cycle times spent by a customer in successive stations are independent in steady state. On the other hand, Burke (1964) considered two single-server queueing systems in tandem with exponential service time and Poisson arrival process and proved that the queue times of a job between consecutive stations are mutually dependent in steady state. However, the above studies on queue time dependence are all derived under Markovian settings.

Although the existence of dependence among stations is well recognized, due to the non-renewal departure processes (Bitran & 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 & 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.

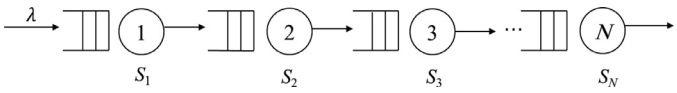

Fig. 1. Tandem queues with N single server stations.

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 Fig. 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 $c_{S_i}^2$, $i = 1, \ldots, N$. Let service rate at station i be μ_i and $\rho_i = \lambda/\mu_i < 1$. The mean queue time at station i is QT_i, $i = 1, \ldots, N$.

Wu and McGinnis (2013) studied tandem queues in Fig. 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 & McGinnis, 2012). 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 $BN_{k+1} = 1$ or 2, then go to step 4. Otherwise, let $k = k + 1$ and go to step 2.
4. If $BN_{k+1} = 2$, then $BN_{k+2} = 1$ and stop. If $BN_{k+1} = 1$, then 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, if the tandem queue in Fig. 1 is an ASIA system, the station i of the tandem queue is a $G/G/1$ queue with the initial arrival process and service time S_i ($1 \le i \le N$) as shown in Fig. 2.