Computers & Industrial Engineering xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

Variability and the fundamental properties of production lines

Kan Wu [a,*], Yaoming Zhou [b], Ning Zhao [c]

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

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Variability
Production lines
Queueing theory
Tandem queues

ABSTRACT

The concept of variability has been commonly used in practice and it is an important performance index of manufacturing systems. In this study, the definition of system variability is given through the insight of Kingman's approximation. The explicit expression for the variability of a production line is derived based on intrinsic ratios and contribution factors. With the derived results, properties of variability for a production line in terms of job arrival rate, service rate and bounds on variability are examined. Simulation results are given to validate the derived properties. The result can be used to guide the design and operations of manufacturing systems.

1. Introduction

The concept of variability is often used by practitioners and researchers to represent the stochastic effect in a manufacturing system. The common sources of variability in manufacturing systems are machine breakdowns, setups, reworks, product mixes, operator availability, batching and fluctuation in process time and arrival intervals (Hopp & Spearman, 2011; Wu, 2014a, 2014b). Reducing variability decreases job queue time, improves system service level (Jacobs, Etman, Van Campen, & Rooda, 2003), and is essential in well-known production control techniques, such as just-in-time (JIT) production (Ohno, 1982), theory of constraints (Goldratt, Cox, & Whitford, 1992) and six sigma (Barney, 2002). Hence, quantifying variability and understanding its basic properties play a key role in achieving effective control of manufacturing systems.

Production managers often want their plants to have higher throughput rate and shorter cycle time under the same capacity. An important question often asked by production managers is that "is our production line more productive than others?" Because it is almost impossible to find two manufacturing systems with the same equipment and capacity, this question is not easy to answer in general. Even if they both have the same type of equipment and capacity, one can have more throughputs but longer cycle time than the other. Then which one is more productive in terms meeting production goals? As we will see in Section 3, the key to answering this question is the variability of manufacturing systems. When a production line is purely deterministic without any randomness, the queue time is zero and the manufacturing system can be operated efficiently in the ideal situation. However, in the presence of randomness, queue time increases at the same utilization. To maintain the same queue time, utilization and thus throughput have to decrease and the return of investment deteriorates. Quantifying and reducing system variability becomes an important theme of manufacturing systems.

In terms of a specific random variable, variability can be simply regarded as the squared coefficient of variation (SCV) (or sometimes the coefficient of variation). Miltenburg (1987) presents a method to determine the asymptotic variance of the output per unit time using the results developed for the asymptotic mean and variance of the total state residence time in Markov chains. Gershwin (1993) presented a method to determine the variance of the output in a given time period from a single station by deriving the difference equations for the probability of producing n parts at a given time and then solving these equations by using some boundary equations. Gershwin also proposed a decomposition method to determine variance of the output from longer lines. Kim and Alden (1997) derived an analytical approximation for the density function and variance of the duration to produce a fixed lot size on a single workstation with deterministic processing times and random downtimes. By using a Markov reward model, Tan (1999) presented a recursive method to determine the mean and variance of the output from a two-station unstable production line with a finite buffer in a given time period conditioned on an arbitrary initial condition. Based on Markovian arrival process, He, Wu, and Li (2007) approximated production variability for a production line with exponential processing times and finite buffers. Both of

* Corresponding author.
E-mail addresses: kan626@gmail.com (K. Wu), zhouyaoming@buaa.edu.cn (Y. Zhou), zhaoning@kmust.edu.cn (N. Zhao).

http://dx.doi.org/10.1016/j.cie.2016.04.014
0360-8352/

the variance of the number of parts produced in a given time period and the variance of the delivery-time to produce a given number of products are discussed.

While reducing variability plays a key role in decreasing system queue time (Delp, Si, & Fowler, 2006), quantifying variability is the first step toward reducing it. Although variability of a random variable (as discussed above) can be rigorously defined by its mean and variance, variability of a production line cannot be defined in such a straightforward manner because: (1) a production line consists of a series of workstations and its performance (i.e., mean sojourn time vs. throughput rate) is the gross effect of a series of operations, and (2) the random variables can be dependent (e.g. the output process), and it is not clear about how to define the variance of a non-renewal process. Due to the non-renewal arrival processes among stations, exact analysis for general production line is not tractable (Berman & Westcott, 1983; Whitt, 1995). On the other hand, due to the nice property of Brownian motion, higher moments other than the first two of the service times and interarrival times have marginal impact on the mean queue time in heavy traffic. When a manufacturing system consists of a single machine with general service times and arrival intervals, the mean queue time of the system can be approximated using Kingman's heavy traffic approximation.

$$QT \leqslant \frac{\sigma_a^2\mu^2 + c_s^2}{2}\frac{\rho}{1-\rho}t \cong \frac{c_a^2 + c_s^2}{2}\frac{\rho}{1-\rho}t = \frac{c_a^2 + c_s^2}{2}QT_{M/M/1}, \tag{1}$$

where QT is the mean queue time of the system, and $QT_{M/M/1}$ is the mean queue time of an $M/M/1$ queue with the same mean arrival rate and service rate as the single machine, c_a^2 is the squared coefficient of variation of arrival intervals, c_s^2 is SCV of service times, μ is service rate, σ_a is the standard deviation of arrival intervals, ρ is utilization and t is the mean service time. In Eq. (1), the first inequality is due to Kingman (1962), and the third term to approximate the queue time is given by Heyman (1975). The service time SCV may come from the small randomness of service time itself, or from the preemptive interruptions as explained by Wu (2014a).

Hopp and Spearman (2011) named the three components of the right-most term in Eq. (1) as VUT, where "V" refers to variability (i.e., $\frac{c_a^2+c_s^2}{2}$), "U" is utilization (i.e., $\frac{\rho}{1-\rho}$), and "T" refers to service time (i.e., t). Hence, system variability (α) of a single machine based on Kingman's approximation is defined as

$$\alpha \equiv \frac{c_a^2 + c_s^2}{2}. \tag{2}$$

While variability of a random variable is characterized by its SCV, variability of a single machine system is characterized through its interarrival time and service time SCV's. The incentive is still to capture the randomness inside a system. Based on Eq. (1), Eq. (2) can be transformed into Eq. (3) as follows (Wu, 2005),

$$\alpha \equiv \frac{QT}{QT_{M/M/1}}. \tag{3}$$

As we will see in Section 3, through the concept of intrinsic ratios, Eq. (3) can be generalized to capture the variability of a general manufacturing system (e.g. a production line) but not limited to a single machine. Rather than defining variability from the ratio of mean and variance (for a random variable), system variability in Eq. (3) is defined based on the ratio of its actual mean queue time to the mean queue time of its corresponding $M/M/1$ queueing system. Although system variability is defined by the ratio of two mean queue times, through Eq. (2), one can see that the fundamentals of system variability still connect to the SCV's of service times and interarrival times.

Although the definition of system variability in Eq. (3) has been given by Wu (2005), the model was based on the stochastic independence assumption (Kleinrock, 1976) which can give large errors in practical situations (Whitt, 1985; Wu & McGinnis, 2013). In this study, we follow the definition of system variability in (Wu, 2005) and investigate its properties under more general settings. The results and insights obtained from the model can be used to guide the activities of managers in manufacturing systems.

This paper is organized as follows. Section 2 reviews intrinsic ratios and queue time approximations. Section 3 explores the properties of system variability. Section 4 validates the models by simulation. Conclusion is given in Section 5.

2. Intrinsic ratio and queue time approximation

To define system variability, we start with production lines consisting of single-server stations as shown in Fig. 1. Assume jobs arrive at the system independently with rate λ. The squared coefficient of variation (SCV) of arrival intervals is c_a^2. There are infinite buffers at each station and the service discipline is first-come first-served (FCFS). Let S_i and $c_{S_i}^2$ be the mean and SCV of the service time at station i. Let service rate at station i be μ_i and $\rho_i = \lambda/\mu_i$. For system stability, assume $\rho_i < 1,\ i = 1,\ldots,N$.

Wu and McGinnis (2013) studied a production line with the structure in Fig. 1 and introduced the concept of intrinsic ratio. Based on intrinsic ratios, an approximate model for the system mean queue time of a general queueing network is derived (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 system mean queue time, both main and sub-bottlenecks of a production line have to be determined first as follows.

Procedure 1 (Identification of bottlenecks)
1. Identify the index of the system bottleneck server (BN_1), where μ_{BN_1} = min μ_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 μ_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 $k = k + 1$, go to 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 production line.

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, station i of the tandem

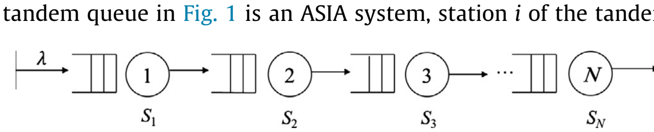

Fig. 1. A production line with N single server stations in series.