Variability and the fundamental properties of production lines

Wu, Kan; Zhou, Yaoming; Zhao, Ning

2016

Wu, K., Zhou, Y., & Zhao, N. (2016). Variability and the fundamental properties of production lines. *Computers & Industrial Engineering*, 99, 364-371.

<https://hdl.handle.net/10356/83989>

<https://doi.org/10.1016/j.cie.2016.04.014>

© 2016 Elsevier. This is the author created version of a work that has been peer reviewed and accepted for publication by *Computers & Industrial Engineering*, Elsevier. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: [<http://dx.doi.org/10.1016/j.cie.2016.04.014>].

Downloaded on 26 Apr 2023 22:46:48 SGT

Variability and the Fundamental Properties of Production Lines

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.

Keywords: variability; production lines; queueing theory; tandem queues

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 et al., 2003), and is essential in well-known production control techniques, such as just-in-time (JIT) production (Ohno, 1982), theory of constraints (Goldratt et al., 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, et al. (2007) approximated production variability for a production line with exponential processing times and finite buffers. Both of 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 et al., 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; Ward 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.

$$QT \leq \frac{\sigma_a^2 \mu^2 + c_s^2}{2} \frac{\rho}{1-\rho} t \cong \frac{c_a^2 + c_s^2}{2} \frac{\rho}{1-\rho} t = \frac{c_a^2 + c_s^2}{2} QT_{M/M/1}, \quad (1)$$

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

$$\alpha \equiv \frac{c_a^2 + c_s^2}{2}. \quad (2)$$

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),

$$\alpha \equiv \frac{QT}{QT_{M/M/1}}. \quad (3)$$

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 (W. 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 Figure 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 $\rho_i < 1$, $i = 1, \dots, N$.

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

Wu and McGinnis (2013) studied a production line with the structure in Figure 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 $\mu_{BN_1} = \min \mu_i$, 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 $k = k + 1$, 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 Figure 1 is an ASIA system, station i of the tandem queue is a $G/G/1$ queue with the initial arrival process and service time S_i ($1 \leq i \leq N$) as shown in Figure 2.

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

Definition 1 In an 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 (J. F. C. Kingman, 1965), the mean queue time (QT_i^A) of station i in an ASIA system can be approximated by

$$QT_i^A \approx \alpha_i \frac{\rho_i}{1 - \rho_i} \frac{1}{\mu_i}, \quad 1 \leq i \leq N. \quad (4)$$

where $\alpha_i = (c_{a1}^2 + c_{S_i}^2)/2$. An ASIA system reduces to a Jackson network (Jackson, 1957) and 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 (or total) mean queue time is purely determined by the bottleneck and is $QT_{BN_k}^A$. It gives the exact solution to a tandem queue (in Figure 1) when all service times are constant (Friedman, 1965). The mean queue time (QT_i^C) of station i in a fully coupled system is

$$QT_i^C = \begin{cases} QT_i, & \text{if } i = 1, \\ QT_i^A - \sum_{k=1}^{i-1} QT_k, & \text{if } i \geq 2 \text{ and station } i \text{ is a bottleneck,} \\ 0, & \text{if } i \geq 2 \text{ and station } i \text{ is not a bottleneck,} \end{cases} \quad (5)$$

where QT_i is the mean queue time at station i in the original tandem queue. 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

$$r_i = \frac{QT_i - QT_i^C}{QT_i^A - QT_i^C}, \quad 2 \leq i \leq N. \quad (6)$$

From simulation experiments, Wu and McGinnis (2013) observed that the intrinsic ratio is approximately linear across traffic intensities. Based on this nice property, Wu and McGinnis (2012) derived Eq. (7) to approximate system mean queue time (QT_f) of N servers in series.

$$QT_f = \sum_{i=1}^N QT_i = \sum_{i=1}^N f_i \cdot QT_i^A \cong \sum_{i=1}^N f_i \cdot \alpha_i \frac{\rho_i}{1 - \rho_i} \frac{1}{\mu_i}, \quad (7)$$

where f_i is called contribution factor (CF) and can be determined by Procedure 2.

Procedure 2 (Determining the parameters f_i)

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

Otherwise, $f_k = r_k * f_k$. Stop if $k = 2$.

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

To validate the approximate errors of Eq. (7) and Procedure 2 in general manufacturing systems, Wu and McGinnis (2012) applied the model to a manufacturing facility in major defense acquisition programs in the US. The assembly line has 14 workstations and 3 operator types with reentrant, rework, batching, various service time distributions and complex product mix. While the largest errors of prior approximate models are larger than 10%, the largest error from Eq. (7) is less than 2%. Hence, although the model is motivated by production lines, it is capable to describe the performance of a general manufacturing system with high fidelity.

3. Variability of a Production Line

In this section we investigate variability of general tandem production lines. Section 3.1 presents the definition of variability of production lines with single server stations. Its properties are studied in Section 3.2. The variability of production lines with multiple-server stations is discussed in Section 3.3.

3.1 Variability of a single-server production line

When a production line consists of single server stations in series, motivated by Eq. (3), variability of the system is defined as follows.

Definition 3 (System variability of a single-server production line)

System variability of a single-server production line is

$$\alpha_f = \frac{QT_f}{QT_{M/M_B/1}}, \quad (8)$$

where QT_f is mean queue time of a production line, $QT_{M/M_B/1}$ is the mean queue time of an $M/M/1$ queue with the same arrival rate and service rate as the system bottleneck of the production line. The system bottleneck of a production line is the station with the highest utilization.

Since system variability is defined based on a single station as expressed in Eq. (3), one may think that we replace the entire production line by a virtual station, and then quantify system variability of the production line from the view point of the virtual station. Because system mean queue time is solely dominated by the system bottleneck (1) in heavy traffic (Iglehart & Whitt, 1970), or (2) when all service times are constant (Friedman, 1965), the system bottleneck station is used to represent the virtual station. Although a production line consist of many stations in series, it is justifiable to compare its performance with a single station, since system queue time can be solely determined by the system bottleneck in some ideal situations. When service time variation departs from zero, system mean queue time becomes longer than the mean queue time of the bottleneck (Wu & McGinnis, 2013) and system variability increases. On the other hand, if all service times are constant, the bottleneck queue time gives a lower bound to the original system. Since Definition 3 is motivated by the reduction method, an entire production line reduces to a single station (i.e., the bottleneck) if all service time variabilities are zero. Furthermore, since the definition is motivated by Kingman's approximation (as shown in Eq. (3)), system variability reduces to the variability of a G/G/1 queue if the system consists of **single stations**.

Throughout the paper, the subscript B is used to denote the system bottleneck. Let $\alpha_B = (c_a^2 + c_{S_B}^2)/2$, $\mu_B = \min_{1 \leq i \leq N} \mu_i$, and $\rho_B = \lambda/\mu_B$. Due to Eq. (7), α_f can be written as

$$\begin{aligned} \alpha_f &= \frac{QT_f}{QT_{M/M_B/1}} \\ &= \alpha_B + \frac{1-\rho_B}{\rho_B} \mu_B \sum_{i \neq B} f_i \alpha_i \frac{\rho_i}{1-\rho_i} \frac{1}{\mu_i} \\ &= \alpha_B + \left(\frac{\mu_B^2}{\lambda} - \mu_B \right) \sum_{i \neq B} f_i \alpha_i \frac{\lambda}{\mu_i - \lambda} \frac{1}{\mu_i} . \end{aligned} \tag{9}$$

Variability of a production line can be decomposed into two parts: the bottleneck variability and the gross effect from the non-bottlenecks, where the contribution from the non-bottlenecks converges to zero

in heavy traffic. That is, system variability is dominated by the bottleneck in heavy traffic, which is consistent with the heavy-traffic bottleneck phenomenon observed by Iglehart and Whitt (1970).

To achieve shorter cycle time with the same throughput rate, maintaining lower variability is desired. Based on Eq. (9), system variability is dominated by the bottleneck variability in heavy traffic. Hence, to prepare for the hot season when system is heavily loaded, production line managers should focus on reducing the variability of system bottleneck. According to Eq. (2), variability can come from poor job dispatching, complex product mix and preemptive interruptions, etc.

3.2 Properties of system variability

If a manufacturing system consists of a single station, its variability behaves like a constant based on Eqs.(1) and (3). It would be interesting to investigate how variability α_f behaves when a manufacturing system consists of multiple single-server stations in series.

(1) Lower variability for higher arrival rate

From Eq. (9), we have

$$\frac{\partial \alpha_f}{\partial \lambda} = \mu_B \sum_{i \neq B} \frac{f_i \alpha_i}{(\mu_i - \lambda)^2} \frac{\mu_B - \mu_i}{\mu_i} < 0.$$

Property 1: Variability of a production line with single-server stations is monotonically decreasing in the external arrival rate.

(2) Lower variability for a small increment of μ_i ($i \neq B$)

Because intrinsic ratios change with μ_i , f_i also depends on μ_i . However, when the value of μ_i does not change much, we can assume f_i in Eq. (9) is (nearly) a constant. In the following, we would like to investigate how system variability changes when the service rate of a non-bottleneck has a small increment (i.e., where the value of f_i does not change much).

Since $\frac{\mu_B}{\lambda} - 1 > 0$, we have

$$\frac{\partial \alpha_f}{\partial \mu_i} = -\mu_B \left(\frac{\mu_B}{\lambda} - 1 \right) f_i \alpha_i \left(\frac{\lambda}{(\mu_i - \lambda)^2 \mu_i} + \frac{\lambda}{(\mu_i - \lambda) \mu_i^2} \right) < 0.$$

Property 2: When the non-bottleneck service rate has a small increment, variability of a production line with single-server stations decreases.

(3) Higher variability for a small increment of μ_B

Following the previous assumption, since $\frac{2\mu_B}{\lambda} - 1 > 0$, we have

$$\frac{\partial \alpha_f}{\partial \mu_B} = \left(\frac{2\mu_B}{\lambda} - 1 \right) \sum_{i \neq B} f_i \alpha_i \frac{\lambda}{\mu_i - \lambda} \frac{1}{\mu_i} > 0.$$

Property 3: When the system bottleneck service rate has a small increment, variability of a production line with single-server stations increases.

(4) Gap effect

Assume F1 and F2 are two almost identical production lines, except that there are small differences in their system bottleneck service rates and non-bottleneck service rates, (so that the value of f_i does not change much). The service rates of system bottleneck and non-bottleneck are μ_{B1} and μ_{L1} in F1, and μ_{B2} and μ_{L2} in F2. The service rate gap is defined as $\mu_{Li} - \mu_{Bi}$, $i = 1, 2$.

According to properties 2 and 3, when μ_{B2} and μ_{L2} satisfy one of the following three conditions, F1 has a lower variability:

- i. $\mu_{L2} < \mu_{L1}$ and $\mu_{B2} = \mu_{B1}$;
- ii. $\mu_{L2} = \mu_{L1}$ and $\mu_{B2} > \mu_{B1}$;
- iii. $\mu_{L2} < \mu_{L1}$ and $\mu_{B2} > \mu_{B1}$.

Property 4: For two almost identical production lines with single-server stations, when one's service rate gap inwardly bounds the other, the one with larger service rate gap has the lower variability.

Remark 1. We often hear from line control managers that a production line with multiple bottlenecks is much harder to control than a production line with a distinct bottleneck. Property 4 indeed confirms this intuition and gives a theoretical foundation for this claim. While a bottleneck is defined based on the average utilization in the long run and is commonly unique, due to stochastic effects, those non-bottlenecks with higher utilizations may also sometimes face congestions and create a transient bottleneck in the short period of time. Hence, when the utilizations of non-bottlenecks are closer to the bottleneck utilization, variability of a production line becomes higher with longer queue time.

Since properties 2, 3 and 4 are derived based on the assumption that f_i is indifferent to the changes of μ_i , they will be validated by simulation in Section 4.

(5) Upper bound of variability

Since $\frac{1-\rho_B}{\rho_B} \frac{\rho_i}{1-\rho_i} \frac{\mu_B}{\mu_i} \leq 1$ in Eq. (9), we have

$$\alpha_f \leq \alpha_B + \sum_{i \neq B} f_i \alpha_i.$$

Property 5: Variability of a production line with single-server stations is upper bounded by $\alpha_B + \sum_{i \neq B} f_i \alpha_i$.

(6) Lower bound of variability

In Eq. (9), $\frac{1-\rho_B}{\rho_B} \mu_B \sum_{i \neq B} f_i \alpha_i \frac{\rho_i}{1-\rho_i} \frac{1}{\mu_i} \geq 0$, hence

$$\alpha_f \geq \alpha_B.$$

Property 6: Variability of a production line with single-server stations is lower bounded by its system bottleneck variability α_B .

(7) Higher variability for a longer production line

Consider a situation that a non-bottleneck station (or station $N + 1$) is added to the production line with N stations (as shown in Figure 1), and $\mu_{N+1} > \mu_B$. Denote the mean queue at station i by QT_i . Under FCFS discipline, the system variability of the production line with $N + 1$ stations is

$$\alpha'_f = \frac{\sum_{i=1}^{N+1} QT_i}{QT_{M/M_B/1}} = \alpha_f + \frac{QT_{N+1}}{QT_{M/M_B/1}}.$$

Therefore, $\alpha'_f > \alpha_f$.

Property 7: When the system bottleneck is unchanged, the variability of a production line with single-server stations increases if more stations are added to the production line.

3.3 Variability of a production line with multiple-server stations

In the previous sections, the production lines with single-server stations are studied. However, a station may have multiple servers in practical manufacturing systems. In the following, we investigate system variability of production lines with multiple-server stations. Assume in a N -station production line, there are m_i servers at station i , $i = 1$ to N . The SCV's of job arrival intervals and S_i are c_a^2 and $c_{S_i}^2$, respectively. Let $\mu_i = 1/E(S_i)$. The system bottleneck of the production line is denoted by B , $m_B\mu_B = \min m_i\mu_i$, for $i = 1$ to N .

When a station is composed of multiple servers, Sakasegawa (1977) proposed the following approximate model based on Kingman's approximation:

$$QT \cong \frac{c_a^2 + c_s^2}{2} \frac{\rho^{\sqrt{2(m+1)}-1}}{1-\rho} \frac{1}{m\mu}, \quad (10)$$

where m is the number of servers at a station and $\rho = \lambda/m\mu$.

Based on Eq. (10), the system mean queue time can be approximated by

$$QT_f \cong \sum_{i=1}^N f_i \alpha_i \frac{\rho_i^{\sqrt{2(m_i+1)}-1}}{1-\rho_i} \frac{1}{m_i \mu_i}, \quad (11)$$

where $\rho_i = \lambda/m_i \mu_i$ is the utilization of station i , $\alpha_i = (c_a^2 + c_s^2)/2$ and f_i is contribution factor of station i .

The mean queue time of system bottleneck with Poisson arrivals and exponential service time is

$$QT_{M/M_B/m_B} = \frac{\rho_B^{\sqrt{2(m_B+1)}-1}}{1-\rho_B} \frac{1}{m_B \mu_B}. \quad (12)$$

From Eq. (11) and Eq. (12), the system variability of a multiple-server production line is

$$\begin{aligned} \alpha_f(m_1, m_2, \dots, m_N) &= \frac{QT_f}{QT_{M/M_B/m_B}} \\ &= \alpha_B + \frac{m_B \mu_B (1 - \rho_B)}{\rho_B^{\sqrt{2(m_B+1)}-1}} \sum_{i \neq B} f_i \alpha_i \frac{\rho_i^{\sqrt{2(m_i+1)}-1}}{1-\rho_i} \frac{1}{m_i \mu_i} \\ &= \alpha_B + \frac{(m_B \mu_B - \lambda)}{\left(\frac{\lambda}{m_B \mu_B}\right)^{\sqrt{2(m_B+1)}-1}} \sum_{i \neq B} f_i \alpha_i \frac{\left(\frac{\lambda}{m_i \mu_i}\right)^{\sqrt{2(m_i+1)}-1}}{m_i \mu_i - \lambda}. \end{aligned} \quad (13)$$

Properties of $\alpha_f(m_1, m_2, \dots, m_N)$ in Eq. (13) can be analyzed through the same way as we did in Section 3.2. When the number of servers at a station is large enough, adding one server to the station won't change the contribution factor f_i much. If f_i in Eq. (11) is (nearly) a constant, we have the following properties.

(1) Higher variability with more servers at the system bottleneck station

If one additional server with service time S_B is added to the system bottleneck, we have

$$\alpha_f(m_1, \dots, m_B + 1, \dots, m_N) - \alpha_f(m_1, \dots, m_B, \dots, m_N)$$

$$= \left[\frac{(m_B \mu_B + \mu_B - \lambda)}{\left(\frac{\lambda}{m_B \mu_B + \mu_B}\right)^{\sqrt{2(m_B+2)}-1}} - \frac{(m_B \mu_B - \lambda)}{\left(\frac{\lambda}{m_B \mu_B}\right)^{\sqrt{2(m_B+1)}-1}} \right] \sum_{i \neq B} f_i \alpha_i \frac{\left(\frac{\lambda}{m_i \mu_i}\right)^{\sqrt{2(m_i+1)}-1}}{m_i \mu_i - \lambda} > 0.$$

Property 8: When the number of servers at the system bottleneck is large, variability of a production line with multiple-server stations is monotonically increasing in the number of servers at the bottleneck.

(2) Lower variability with more servers at a non-bottleneck station ($k \neq B$)

If one additional server with service time S_k is added to station k , we have

$$\begin{aligned} & \alpha_f(m_1, \dots, m_k + 1, \dots, m_N) - \alpha_f(m_1, \dots, m_k, \dots, m_N) \\ &= \frac{(m_B \mu_B - \lambda)}{\left(\frac{\lambda}{m_B \mu_B}\right)^{\sqrt{2(m_B+1)}-1}} \left[f_k \alpha_k \frac{\left(\frac{\lambda}{m_k \mu_k + \mu_k}\right)^{\sqrt{2(m_k+2)}-1}}{m_k \mu_k + \mu_k - \lambda} - f_k \alpha_k \frac{\left(\frac{\lambda}{m_k \mu_k}\right)^{\sqrt{2(m_k+1)}-1}}{m_k \mu_k - \lambda} \right] < 0. \end{aligned}$$

Property 9: When the number of servers at a non-bottleneck station is large, variability of a production line with multiple-server stations is monotonically decreasing in the number of servers at the non-bottleneck.

Properties 8 and 9 are consistent with the gap effect in Property 4. When more servers are added to the bottleneck station, the bottleneck utilization decreases. The gap becomes smaller and system variability increases. When more servers are added to a non-bottleneck, the non-bottleneck utilization decreases. The gap becomes larger and system variability decreases.

(3) Lower variability with more servers at each station

Consider two production lines with the same type of machines and making the same type of product. Let the utilizations of two production lines be the same, so that one of them has k times more servers at each

station and k times more input at the first station ($k > 1$). When m_i is large and k is close to 1, f_i in Eq. (9) can be assumed to be a constant. In this situation, we would like to compare the variability of two production lines.

From Eq. (13), the system variability with m_i servers at station i ($i = 1, \dots, N$) is

$$\begin{aligned} & \alpha_f(m_1, m_2, \dots, m_N) \\ &= \alpha_B + \sum_{i \neq B} f_i \alpha_i \frac{(m_B \mu_B - \lambda) \rho_i^{\sqrt{2(m_i+1)}-1}}{(m_i \mu_i - \lambda) \rho_B^{\sqrt{2(m_B+1)}-1}}. \end{aligned} \quad (14)$$

The system variability with km_i servers at station i ($i = 1, \dots, N$) is

$$\begin{aligned} & \alpha_f(km_1, \dots, km_k, \dots, km_N) \\ &= \alpha_B + \frac{(km_B \mu_B - k\lambda)}{\left(\frac{k\lambda}{km_B \mu_B}\right)^{\sqrt{2(km_B+1)}-1}} \sum_{i \neq B} f_i \alpha_i \frac{\left(\frac{k\lambda}{km_i \mu_i}\right)^{\sqrt{2(km_i+1)}-1}}{km_i \mu_i - k\lambda} \\ &= \alpha_B + \sum_{i \neq B} f_i \alpha_i \frac{(m_B \mu_B - \lambda) \rho_i^{\sqrt{2(km_i+1)}-1}}{(m_i \mu_i - \lambda) \rho_B^{\sqrt{2(km_B+1)}-1}}. \end{aligned} \quad (15)$$

Let $X_i = f_i \alpha_i \frac{(m_B \mu_B - \lambda) \rho_i^{\sqrt{2(m_i+1)}-1}}{(m_i \mu_i - \lambda) \rho_B^{\sqrt{2(m_B+1)}-1}}$ in Eq. (14) and $Y_i = f_i \alpha_i \frac{(m_B \mu_B - \lambda) \rho_i^{\sqrt{2(km_i+1)}-1}}{(m_i \mu_i - \lambda) \rho_B^{\sqrt{2(km_B+1)}-1}}$ in Eq. (15). If

$$m_i > m_B,$$

$$\frac{Y_i}{X_i} = \frac{\rho_i^{\sqrt{2(km_i+1)}-1}}{\rho_B^{\sqrt{2(km_B+1)}-1}} \leq \left(\frac{\rho_i}{\rho_B}\right)^{\sqrt{2(km_B+1)}-1} < 1.$$

Comparing the terms X_i and Y_i in Eq. (14) and Eq. (15), we have

$$\alpha_f(km_1, \dots, km_k, \dots, km_N) < \alpha_f(m_1, m_2, \dots, m_N).$$

Property 10: When there are many servers at each station and the system bottleneck has the least number of servers among all stations, if a production line has more servers than the other at each station, the variability of the production line is smaller.

In order to obtain higher throughput rate with the same investment, the bottleneck of a production line tends to be allocated to the most expensive machine with less number of servers. In this situation, based on Property 10, to enjoy the economics of scale, a production line designed to produce large volume will have lower variability and is easier to manage.

Remark 2. In the semiconductor industry, economics of scale is much emphasized. The capacity of a fab has become larger and larger from a couple thousands wafers per month in the 80s in the US to over 100k wafers per month nowadays in Asia. The clean room area grows dramatically and the tool group becomes much larger than before. Property 10 gives a quantitative measure and partially explains this phenomenon in the long term production planning.

Since the other properties are similar to the single-server cases, we omit them here.

4. Simulation Validation

Since Eq. (7) is only an approximation for mean queue time, it would introduce error to α_f . In the following we present simulation results to examine the properties of variability for production lines with single-server stations.

First, we consider three different production lines with three stations each and the three production lines only differ in S_1 , where $S_1 \sim \text{Gamma}(5/3, 15)$, $\text{Gamma}(5/3, 9)$ and $\text{Gamma}(5/3, 3)$ respectively. Hence, $\mu_1 = 1/25$, $1/15$ and $1/5$ respectively. Let $S_2 \sim \text{Gamma}(5/2, 12)$, $S_3 \sim \text{Gamma}(5/4, 112/5)$. Therefore, $\mu_2 = 1/30$, $\mu_3 = 1/28$. Therefore, station 2 is the system bottleneck with $\alpha_B = 0.7$. Arrivals follow the Poisson distribution. Job arrival rate is $\lambda = \rho\mu_2$, where ρ varies from 0.1 to 0.95. 30 replications are conducted at each utilization. Each of the 30 replications is composed of 200,000 jobs after discarding the first 400,000 jobs for warm-up. The simulation mean queue times (QT_f^S) followed by its half-width of the 95% confidence intervals, and the system variabilities are presented in Table 1. α_f^S is

the system variability computed based on simulation queue time, $\alpha_f^S = \frac{QT_f^S}{QT(M/M_B/1)}$. α_f is the system variability computed based on approximation queue time, $\alpha_f = \frac{QT_f}{QT(M/M_B/1)}$. Since the difference between α_f^S and α_f is small, our approximation of system variability (α_f) based on Eq. (7) is accurate. Furthermore, since $\alpha_B = 0.7$, $\alpha_f > 0.7$ in all cases as expected.

Table 1. Mean queue times and variabilities with respect to non-bottleneck service rates

ρ	$\mu_1=1/25$			$\mu_1=1/15$			$\mu_1=1/5$		
	QT_f^S	α_f^S	α_f	QT_f^S	α_f^S	α_f	QT_f^S	α_f^S	α_f
0.10	5.617±0.020	1.685	1.683	4.615±0.021	1.384	1.383	4.162±0.021	1.249	1.249
0.20	12.498±0.046	1.666	1.666	10.264±0.034	1.369	1.369	9.350±0.033	1.247	1.247
0.30	21.142±0.088	1.644	1.646	17.395±0.051	1.353	1.352	15.912±0.058	1.238	1.238
0.40	32.470±0.103	1.624	1.623	26.679±0.110	1.335	1.336	24.692±0.077	1.235	1.235
0.50	47.731±0.164	1.591	1.593	39.535±0.161	1.318	1.317	36.783±0.134	1.226	1.226
0.60	70.288±0.283	1.562	1.561	58.451±0.267	1.299	1.299	55.050±0.216	1.223	1.223
0.70	106.369±0.613	1.520	1.519	89.273±0.533	1.275	1.276	85.310±0.176	1.219	1.214
0.80	175.215±1.573	1.460	1.461	151.079±1.028	1.259	1.259	145.668±0.457	1.214	1.223
0.90	366.286±5.620	1.357	1.359	335.712±4.462	1.243	1.243	328.492±1.631	1.211	1.211
0.95	733.621±23.005	1.287	1.287	704.989±21.629	1.237	1.237	680.365±19.282	1.194	1.160

Figure 3. Variability of three production lines with different non-bottleneck service rates

The trends in system variabilities α_f with respect to different job arrival rates for the three cases are compared and presented in Figure 3. Figure 3 shows that variability of production lines is monotonically decreasing in its external arrival rate and is also decreasing in the non-bottleneck service rate as we expected. Because system queue time is dominated by the system bottleneck in heavy traffic, system variabilities of the three systems differ more in light traffic and less in heavy traffic.

To investigate property 3, another three production lines are compared. Assume there are three stations in each production line and the three production lines only differ in the system bottleneck service time S_2 , where $S_2 \sim \text{Gamma}(5/2, 16)$, $\text{Gamma}(5/2, 14)$, or $\text{Gamma}(5/2, 12)$ respectively. Hence, $\mu_2 = 1/40$, $1/35$ and $1/30$. Let $S_1 \sim \text{Gamma}(5/3, 15)$ and $S_3 \sim \text{Gamma}(5/4, 112/5)$ (where $\mu_1 = 1/25$ and $\mu_3 = 1/28$). Therefore, station 2 is the bottleneck with $\alpha_B = 0.7$. Arrivals follow the Poisson distribution. Job arrival rate is $\lambda = \rho/40$, where ρ varies from 0.1 to 0.95. 30 replications are conducted at each utilization. Each of the 30 replications is composed of 200,000 jobs after discarding the first 400,000 jobs for warm-up. The results are given in Table 2, where $\alpha_f^S = \frac{QT_f^S}{QT(M/M_B/1)}$, and $\alpha_f = \frac{QT_f}{QT(M/M_B/1)}$. In Table 2, one can observe that α_f is a good approximation of α_f^S , and $\alpha_f > \alpha_B$.

Table 2. Table 2. Mean queue times and variabilities with respect to system bottleneck service rates

λ	$\mu_2=1/40$			$\mu_2=1/35$			$\mu_2=1/30$		
	QT_f^S	α_f^S	α_f	QT_f^S	α_f^S	α_f	QT_f^S	α_f^S	α_f
0.0025	6.554±0.028	1.475	1.474	5.222±0.024	1.556	1.556	4.100±0.020	1.685	1.688
0.0050	14.542±0.046	1.454	1.455	11.437±0.035	1.540	1.540	8.868±0.032	1.675	1.676
0.0075	24.632±0.092	1.437	1.437	18.943±0.062	1.521	1.521	14.470±0.048	1.661	1.663
0.0100	37.735±0.132	1.415	1.414	28.340±0.106	1.504	1.504	21.152±0.078	1.645	1.645
0.0125	55.688±0.218	1.392	1.393	40.261±0.131	1.479	1.480	29.245±0.107	1.625	1.626
0.0150	81.978±0.392	1.366	1.366	56.276±0.207	1.455	1.456	39.472±0.179	1.608	1.609
0.0175	123.850±0.656	1.327	1.328	78.985±0.325	1.428	1.427	52.560±0.207	1.585	1.584
0.0200	207.701±1.589	1.298	1.298	113.990±0.604	1.396	1.396	70.374±0.348	1.564	1.562
0.0225	449.668±5.442	1.249	1.249	176.900±1.187	1.364	1.364	95.172±0.491	1.527	1.527
0.0237	937.063±10.288	1.233	1.233	228.533±2.075	1.326	1.327	112.576±0.640	1.514	1.514

The trends in system variabilities α_f with respect to different job arrival rates of the three cases are compared and presented in Figure 4. Figure 4 shows that the variability of production lines is monotonically decreasing in its external arrival rate and is increasing in the service rate of system bottleneck. When $\mu_2 = 1/40$ and $1/30$, the service rate gaps are $3/200$ (or $1/25 - 1/40$) and $1/150$ (or $1/25 - 1/30$) respectively. Figure 4 shows that α_f is smaller when $\mu_2 = 1/40$. Therefore, for two identical production lines with single-server stations, when one service rate gap inwardly bounds the other, the one with larger service rate gap has the lower variability as stated in Property 4.

Figure 4. Variability of three production lines with different system bottleneck service rates

5. Conclusion

The concept of system variability is commonly used to delineate the stochastic effects in manufacturing systems. In this study, we clearly define this terminology. Different from the definition of variability for a random variable, system variability is defined by the mean queue times inspired by Kingman's approximation.

Through the concepts of intrinsic ratios and contribution factors, properties of system variability are investigated. In addition to service time and interarrival time variabilities, we find that system variability will increase if external arrival rate decreases, non-bottleneck service rate decreases, system bottleneck service rate increases or production line becomes longer. When two systems have different system variabilities, the one with the smaller one has higher productivity in terms of its queue time performance. System variability is an important performance index, and can be used to compare the performance of different manufacturing systems more objectively. The proposed model and results can also be applied to many areas in the field of manufacturing management, such as system performance analysis, capacity planning, and queue time reduction.

Based on the theory of constraints, in order to increase the capacity of a production line, a production manager should try to increase the capacity of a bottleneck until all bottlenecks are relieved. However, based on Property 4, a stochastic production line without a unique bottleneck is more difficult to manage. Indeed, if a production line has no bottleneck, then every station is a bottleneck. Hence, productivity improvement should not only focus on the first moment (i.e., capacity), but also consider the higher moments, e.g. the variance of service times. Wu and Zhao (2015) have found that the most effective place to reduce system cycle time may not be the throughput bottleneck station.

The exact relations among intrinsic ratios, service time and interarrival time variabilities, and service rate are still not clear. Although the properties perform well with different values of service time in simulation, f_i in Eqs. (9) and (13) depends on μ_i , and Properties 2, 3 and 4 are valid only for a small change of service times in our proofs. Those properties can be applied to a more general situation, if all the relations become clear. This study of intrinsic ratios is left to future research.

Reference

- Barney, M. (2002). Motorola's second generation. Six Sigma Forum Magazine, 1(3), 13-16.
- Berman, M. & Westcott, M. (1983). On queueing systems with renewal departure processes. Advances in Applied Probability, 657-673.

- Delp, D., Si, J. & Fowler, J.W. (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.
- Gershwin, S. (1993). Variance of output of a tandem production system. In: R. Onvural & I. Akyldiz, *Second International Workshop on Queuing Networks with Finite Capacity* (pp. 291-304). North Carolina: Elsevier.
- Goldratt, E.M., Cox, J. & Whitford, D. (1992). *The goal: a process of ongoing improvement* (Vol. 2): North River Press Great Barrington, MA.
- He, X.-F., Wu, S. & Li, Q.-L. (2007). Production variability of production lines. *International Journal of Production Economics*, 107(1), 78-87.
- Heyman, D.P. (1975). A diffusion model approximation for the GI/G/1 queue in heavy traffic. *Bell System Tech. J*, 54, 1637-1646.
- Hopp, W.J. & Spearman, M.L. (2011). *Factory physics*: Waveland Press.
- Iglehart, D.L. & Whitt, W. (1970). Multiple channel queues in heavy traffic. II: Sequences, networks, and batches. *Advances in Applied Probability*, 2(2), 355-369.
- Jackson, J.R. (1957). Networks of waiting lines. *Operations research*, 5(4), 518-521.
- Jacobs, J., Etman, L., Van Campen, E. & Rooda, J. (2003). Characterization of operational time variability using effective process times. *Semiconductor Manufacturing, IEEE Transactions on*, 16(3), 511-520.
- Kim, D. & Alden, J. (1997). Estimating the distribution and variance of time to produce a fixed lot size given deterministic processing times and random downtimes. *International Journal of Production Research*, 35(12), 3405-3414.
- Kingman, J. (1962). Some inequalities for the queue GI/G/1. *Biometrika*, 49(3/4), 315-324.
- Kingman, J.F.C. (1965). The heavy traffic approximation in the theory of queues. In: W.L. Smith & W.E. Wilkinson, *Symposium on Congestion Theory* (pp. 137-159). University of North Carolina at Chapel Hill: University of North Carolina Press.
- Kleinrock, L. (1976). *Queueing Systems: Volume 2: Computer Applications* (Vol. 82): John Wiley & Sons.
- Miltenburg, G. (1987). Variance of the number of units produced on a transfer line with buffer inventories during a period of length T. *Naval Research Logistics (NRL)*, 34(6), 811-822.
- Ohno, T. (1982). How the Toyota production system was created. *Japanese Economy*, 10(4), 83-101.

- Sakasegawa, H. (1977). An approximation formula $L_q = \alpha \cdot \rho^\beta / (1 - \rho)$. *Annals of the Institute of Statistical Mathematics*, 29(1), 67-75.
- Tan, B. (1999). Variance of the output as a function of time: Production line dynamics. *European Journal of Operational Research*, 117(3), 470-484.
- Whitt, W. (1985). The best order for queues in series. *Management Science*, 31(4), 475-487.
- 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. *Semiconductor Manufacturing, IEEE Transactions on*, 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. & McGinnis, L. (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. & McGinnis, L. (2013). Interpolation approximations for queues in series. *IIE transactions*, 45(3), 273-290.
- Wu, K. & Zhao, N. (2015). Dependence among single stations in series and its applications in productivity improvement. *European Journal of Operational Research*, 247(1), 245-258.