An examination of variability and its basic properties for a factory

Wu, Kan

2005

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

<https://hdl.handle.net/10356/94253>

<https://doi.org/10.1109/TSM.2004.840525>

© 2005 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [DOI: <http://dx.doi.org/10.1109/TSM.2004.840525>].

Downloaded on 26 Apr 2023 11:30:58 SGT

An Examination of Variability and Its Basic Properties for a Factory

Kan Wu

*IE Department, Inotera Memories Inc., Taoyuan, Taiwan, R.O.C.
(e-mail: kanwu@inotera.com)*

Abstract

Variability is a key performance index of a factory. In order to characterize variability of a factory, definitions of bottleneck, utilization, and variability of a single machine are reexamined and clarified. The clarification leads to the introduction of a detail expression for the relationship between cycle time and work-in-progress.

In order to quantify variability for factories, the author uses a single machine system to gauge the behaviors, and subsequently derives an explicit expression for the variability, of a simple factory, making use of analogy and the clarified definitions. The obtained results can be applied to many subjects in the field of manufacturing management, such as factory performance analysis, capacity planning, and cycle time reduction.

With the derived results, properties of variability for a simple factory in the aspects of utilization versus throughput bottlenecks and nonthroughput bottlenecks, gap effects, and bounds on variability, are examined in detail to shed light on the insights of the stochastic behaviors of a complex factory.

Index Terms—Bottleneck, performance, utilization, variability.

I. INTRODUCTION

ACCORDING to the International Technology Roadmap for Semiconductors (ITRS) [1], the average capital investment for a 300-mm semiconductor wafer fab can be as high as US\$2.7 billion, of which, US\$2.16 billion (80%) will be spent on production equipment. In order to effectively operate a costly fab, a “cost/function” improvement of about 20%–30% is needed each year. Such improvement can be realized from decreases in device feature sizes, wafer size increases, improvements of yield and Overall Equipment Efficiency (OEE). Since contributions from wafer size increases and yield improvements have slowed down since 1995, the relative importance of OEE improvement has grown significantly and is expected to maintain a 9%–15% improvement per year in order to preserve the “cost/function” improvement rate.

One direct way to improve OEE is to increase utilization. In a stochastic production environment, higher utilization can be achieved by reducing variability, or by increasing cycle time. However, increasing cycle time may cause unfavorable yields, market loss, or unsatisfied customers. Furthermore, increasing cycle time will inevitably increase the work-in-progress (WIP) level of a factory. Therefore, cycle times cannot be increased without limit.

Another way to achieve higher utilization is to reduce variability. Variability is caused by random events within a fab, such as machine failures, setup, batching rules, and rework. Some of these are controllable, such as setup, batching rules, and rework; some can be classified as pure randomness, such as machine failures. Achieving lower variability is always preferred since it leads to higher utilization without increasing cycle times. Therefore, reducing variability is treated as a productivity improvement objective [2]–[11].

Quantifying variability is the first step toward reducing it. Variability can be characterized by both queueing time and utilization, as illustrated in Fig. 1 [12]. Examinations of variability for a single machine or a single type of machine have been fully covered [12]–[20]. However, examination of variability for factory is more complicated. Apart from simulation studies [2], [4], [8], [10], [21], [22], Boebel and Halmel attempted to identify the characteristic curve of a factory based on the G/G/1 queueing model and demonstrated how to enhance productivity through variability reduction [23]. However, instead of a single server, a factory is composed of a sequence of operations executed by a series of servers. Since factory variability is the gross effect of the process time variability and flow variability of each machine, predicting its characteristic curve simply based on the G/G/1 queue is not a suitable approach.

In a semiconductor fab, the manufacturing environment is highly stochastic due to the combined effects of the following occurrences: rework, scraps, reentry, different lot due dates, setup, batching, heuristic dispatching rules, operator availability, preventative maintenance, and machine breakdown. Applications of queueing network models to assess performance of complex semiconductor fabs on system levels have been developed [24], [25].

Yet, the property of system variability is not fully examined. This paper will first define the basic terminologies of a factory, and then derive its system variability. Properties of a factory are subsequently examined by means of the derived results.

II. GENERAL DEFINITIONS AND VARIABILITY OF A SINGLE MACHINE

A. Definition of Bottlenecks

Previous studies have tried to characterize properties of a bottleneck. Wein selected stations that are heavily utilized (approximately 90% utilization) as the fab bottlenecks [26]. Lozinski and Glassey defined the bottleneck machine to be the machine with the smallest idle time [27]. Goldratt and Cox described the bottleneck as the machine with over 100% utilization [28]. Martin described the bottleneck as the toolset with the highest X-factor [29]. In applications on real shop floors, the bottleneck is sometimes defined as the machine with the highest queueing WIP level or the longest lot queueing time.

In this paper, we define bottlenecks as *a bottleneck is the constraint that prevents a factory from attaining its production goal(s)*. Therefore, different kinds of bottlenecks may exist according to different production goals. A throughput bottleneck prevents a factory from achieving a higher throughput rate, while a cycle time bottleneck prevents it from achieving shorter total cycle time.

If the bottleneck type is treated as a throughput bottleneck, it can be defined as, *when demand increases continuously, the first machine whose utilization exceeds 100%*. In other words, if the throughput rate of a factory needs to be increased, the capacity of the throughput bottleneck needs to be expanded. If the machine with the highest utilization will be the first machine whose utilization exceeds 100% when demand increases, a throughput bottleneck can be predicted as *the machine with the highest utilization*. Therefore, although a throughput bottleneck theoretically always exists, it may not always be active. When utilization of all machines is far below 100%, no active throughput bottleneck exists at that moment.

While a throughput bottleneck is a specific machine or group of machines, not all bottlenecks can be attributed to a specific machine group. A cycle time bottleneck prevents a factory from achieving shorter total cycle time. For any lot, the total cycle time is the sum of the average cycle time per visit of every machine on the critical path times the reentry frequency (r_i) at that machine.

$$\text{Total Cycle Time} = \sum_i CT_i * r_i \quad (1)$$

where i is the visited machine on the critical path, and CT_i is the average cycle time per visit of machine i . For a semiconductor fab, critical path is the same as the main process route. Since the cycle time of any machine on the critical path contributes to the total cycle time, each machine forms part of the cycle time bottleneck with differing degrees of seriousness.

However, a throughput bottleneck is possibly the most serious cycle time bottleneck. The approximation for queueing time (QT) of a single machine, which was first investigated by Kingman in 1961, is given by

$$QT(G/G/1) = \left(\frac{C_a^2 + C_e^2}{2} \right) \left(\frac{u}{1-u} \right) EPT \quad (2)$$

where u is the utilization, EPT is the mean effective process time, C_a^2 is the squared-coefficient-of-variation (SCV) of the arrival interval, and C_e^2 is the SCV of the effective process time [15]. SCV is " σ^2/μ^2 ," where σ^2 is the variance, and μ is the mean. Here, EPT of a single machine is defined as

$$EPT = \frac{\text{Utilization}}{\text{Throughput Rate}} \quad (3)$$

According to Sattler, EPT of a single machine can be realized as all cycle time except waiting for another lot [2]. EPT can also be realized as "the total amount of time a lot claims capacity of a machine, even if it is not yet being processed" [30]. It should include the raw process time and all time losses due to setup, breakdown, unavailability of operator, and any other source of variability. Cycle time is defined as

$$CT = PT + QT \quad (4)$$

where PT is the raw process time. The raw process time should include all the time periods that a lot spends on a machine, such as load, orientation, pumping down, robot transferring, processing, and unload. Although the total PT for a series of operations is the summation of all PT at every operation, later we will show that system EPT is not simply the summation of all EPT at every operation.

Since cycle time increases rapidly when utilization is high, if the SCVs for each machine and process times are similar, higher utilization shall lead to much longer cycle time. This approach differs from the thoughts of previous studies in that, in a stochastic environment, factory bottlenecks mainly come from cycle time rather than from throughput constraints. When demand exceeds capacity, factory throughput is required to be as much as possible. However, cycle time cannot be increased without limit. Since cycle time increases rapidly when utilization is high, it already becomes a constraint well before utilization reaches 100%.

The overall cycle time is the summation of cycle times from all visited machines on the critical path. Under steady-state conditions and for a fixed target of total cycle time, reducing the cycle time of one machine (by improving its variability or process time) allows cycle time increases for other machines. If increase of cycle time of other machines is allowed and the variability of each machine is unchanged, throughput may be increased, without violating the target of total cycle time. The more the cycle time is reduced for that machine, the higher the factory throughput that can be achieved.

B. Definition of Utilization

Many types of utilization have been defined according to different situations. SEMI Standards E10 defines “operational utilization” as “productive time/operations time” and “total utilization” as “productive time/total time” [31], where operations time is the total time excluding nonscheduled time, and productive time includes regular production, engineering runs, rework and work for third party. Utilization is also defined as “production process time/tool production available time” [11], [18], where production process time is the total PT , and tool production available time is the available time for product production operations, which is similar to the “manufacturing time” defined in E10.

However, following the terminology of queueing theory [12], utilization (u) used in (2) should be defined as “arrival rate/service rate”, which also can be expressed as “ EPT /mean inter-arrival time.” Defining general production time (GPT) as the total time except all the time waiting for lots, or as the summation of all during the total observation period (i.e., total time). Utilization can be expressed as

$$\text{Utilization} = \frac{GPT}{\text{Total Time}} \quad (5)$$

where GPT equals “total time * EPT /mean inter-arrival time.” This definition is different from the ones mentioned previously, but is the one that should be applied to the queueing model.

Under steady-state conditions, arrival rate is the same as actual throughput rate. Utilization is given by

$$\text{Utilization} = \frac{\text{Actual Throughput Rate}}{\text{Service Rate}}. \quad (5a)$$

For a system composed of a series of machines, under steady-state conditions, the actual system throughput rate is the same as any machine type's throughput rate divided by the reentry frequency at the machine type. Furthermore, the system service rate is constrained by the machine service rate of the throughput bottleneck (TB) divided by the reentry frequency at the bottleneck. Therefore, system utilization is given by

$$\begin{aligned} \text{System Utilization} &= \frac{\frac{\text{Any Throughput Rate}}{\text{Its Re-entry Freq.}}}{\frac{\text{TB Service Rate}}{\text{TB Re-entry Freq.}}} \\ &= \frac{\frac{\text{TB Throughput Rate}}{\text{TB Re-entry Freq.}}}{\frac{\text{TB Service Rate}}{\text{TB Re-entry Freq.}}} \\ &= \text{Throughput Bottleneck Utilization}. \end{aligned} \quad (6)$$

Definition 1: Under steady-state conditions, system utilization equals the throughput bottleneck utilization.

C. Variability of a Single Machine

Machine variability is caused by the random nature of a production system. The causes can be classified into two categories: process time variability and flow variability [15]. Process time variability refers to machine breakdown, setup, rework, product mix, operator availability, and natural fluctuation in process time. Flow variability refers to the varying arrival intervals due to upstream station process time variabilities. Between any two consecutive process steps, the upstream station's departure SCV becomes the downstream station's arrival SCV. According to Buzacott and Shanthikumar [14], the departure SCV can be estimated as follows:

$$C_d^2 = 1 + (1 - u^2)(C_a^2 - 1) + \left(\frac{u^2}{\sqrt{m}}\right)(C_e^2 - 1) \quad (7)$$

where m is the number of machines at the upstream station, and C_a^2 is the upstream station's arrival SCV. From (7), the flow variability of the downstream station is affected by the utilization of the upstream stations. Both process time variability and flow variability constitute the variability in (2), where the system variability (α) of a single machine is approximated by

$$\alpha = \frac{C_a^2 + C_e^2}{2} \quad (8)$$

Two important implicit assumptions of the G/G/1 queue are that process time variability is independent of flow variability, and variability is independent of current machine utilization. Based on (2), the variability (α) of a single machine can be re-expressed by the following:

$$\alpha = \frac{QT}{\left(\frac{u}{1-u}\right) EPT}. \quad (9)$$

Observation 1: If the process time variability and flow variability of a machine during two different time periods are the same, its variability is the same during those two time periods as well.

Note that Observation 1 does not require utilization to be the same for those two periods. Based on (2), if cycle time, process time, and utilization are known, variability can be calculated. However, in practice, if the observation duration is not long enough, utilization is difficult to measure objectively, since it changes intermittently over time, and is upper-bounded by one.

To overcome this issue, another form of (2) was derived, relying on inventory information (WIP) rather than utilization. Compared to utilization, inventory is more observable and has no boundary, therefore, easier to be measured objectively. For a single machine, based on (2), a revised X-factor (X) is defined by the following:

$$X = \frac{CT}{PT} = \alpha \frac{u}{1-u} \frac{EPT}{PT} + 1. \quad (10)$$

Note that (10) is different from previous studies [11], [16]–[18], [23], since “ EPT/PT ” may not equal one. Little’s Law [32] in queueing theory states that

$$W = \lambda * CT \quad (11)$$

where W is inventory, λ is the mean lot arrival rate, and CT is the cycle time. By introducing Little’s Law to (2), (10) can be transformed into a more useful form.

$$X = \frac{1}{2} \left[(1 + \beta) + \sqrt{\beta^2 + 2 \left(2\alpha \frac{EPT}{PT} - 1 \right) \beta + 1} \right] \\ \text{and } \beta = \frac{W}{PT \times CP} \quad (12)$$

where W is inventory, and CP is capacity.

From (12), variability (α) can be obtained as

$$\alpha = \frac{(X - 1)(X - \beta)}{\beta} \frac{PT}{EPT} \quad (12a)$$

without requiring any utilization information. An important relationship is observed from (12): as WIP level increases, cycle time approaches an asymptote with *slope* = $1/CP$ and *intercept* = $\alpha \cdot EPT$ as illustrated in Fig. 2. According to (10) and (12), utilization, cycle time, and WIP are dependent on each other. When the capacity and characteristic curve are fixed during a certain period of time, knowing any one of the above three parameters leads to the determination of the other two.

III. VARIABILITY OF A SIMPLE FACTORY

A simple factory is defined as a series of operations processed by a series of machines. Each machine follows the behaviors of G/G/1 queue. Furthermore, each machine is dedicated to at least one operation (i.e. reentry is acceptable) and each operation corresponds to one machine only. Machine process time variabilities are independent of flow variabilities. Queueing WIP is allowed at each machine.

Factory variability is the gross effect of the process time variability and flow variability of each machine. Since different queueing WIP profiles imply different utilization distributions, which lead to different flow variability distributions according to (7), factory variability is closely related to the internal utilization distribution. When process time variability is fixed, the only way to guarantee that flow variability at each machine is the same during two different time periods is to ensure that the utilization of each machine (i.e., utilization distribution) is the same. If Observation 1 is extended to a factory, it needs to be limited to a specific WIP profile or utilization distribution.

Observation 1A: Inside a factory, if the process time variability and flow variability of each machine during two different time periods are the same, factory variability is the same during those two time periods only if their utilization distributions during those two time periods are the same.

Assume there exists a single machine system whose system utilization (u_s), effective process time (EPT_s), and queueing time (QT_s) are identical to a simple factory system of n operations and m machines ($n \geq m$) at a specific utilization distribution, $\hat{u}_s$, as demonstrated in Fig. 3. According to Observation 1A, the variability of a simple factory system (α_f) can be gauged by the variability of a single machine system (α_s) at a specific utilization distribution, when their utilization, effective process times, and queueing times are the same.

Since QT_s equals QT_f at $\hat{u}_s$, based on the assumptions, the (QT_f) factory queueing time is expressed as the following:

$$\begin{aligned} QT_f|_{at \hat{u}_s} &= QT_s(G/G/1) = \alpha_s \left(\frac{u_s}{1 - u_s} \right) EPT_s \\ &= \alpha_f \left(\frac{u_f}{1 - u_f} \right) EPT_f|_{at \hat{u}_s} \end{aligned} \quad (13)$$

where α_f is the variability of a simple factory, EPT_f is the mean effective process time of a simple factory, u_f is the utilization of a simple factory, and $\hat{u}_s = (u_1, u_2, \dots, u_m)$.

Based on (3), for a simple factory system, the TB effective process time (EPT_{TB}) and EPT_f are given by

$$EPT_{TB} = \frac{TB \text{ Utilization}}{TB \text{ Throughput Rate}} \quad (14a)$$

$$EPT_f = \frac{\text{System Utilization}}{\text{System Throughput Rate}}. \quad (14b)$$

Under steady-state conditions, system throughput rate is the same as the TB throughput rate divided by the TB reentry frequency (r_{TB}), and, from Definition 1, system utilization (u_f) is equal to the TB utilization (u_{TB}), therefore,

$$EPT_f = EPT_{TB} * r_{TB}. \quad (14c)$$

Based on the assumptions and (14c), (13) becomes

$$QT_f|_{at \hat{u}_s} = \alpha_f \left(\frac{u_{TB}}{1 - u_{TB}} \right) EPT_{TB} * r_{TB}|_{at \hat{u}_s}. \quad (15)$$

For any given lot, its total cycle time is the sum of the average cycle time per visit of every machine times the reentry frequency (r_i) as (1). X-factor for the factory (X_f) is given by

$$\begin{aligned} X_f &= \frac{CT_f}{PT_f} \\ &= \frac{r_1 CT_1 + r_2 CT_2 + \dots + r_m CT_m}{PT_f} \\ &= \frac{PT_1}{PT_f} \frac{r_1 CT_1}{PT_1} + \frac{PT_2}{PT_f} \frac{r_2 CT_2}{PT_2} + \dots + \frac{PT_m}{PT_f} \frac{r_m CT_m}{PT_m} \\ &= r_1 f_1 X_1 + r_2 f_2 X_2 + \dots + r_m f_m X_m \\ &= \frac{QT_f}{PT_f} + 1 \end{aligned} \quad (16)$$

where CT_f is the total cycle time, PT_f is the total process time, PT_i is the process time of machine i , and $f_i = PT_i/PT_f$.

Based on (15) and (16), at any given utilization distribution, $\hat{u}_s$, QT_f is given by

$$QT_f|_{at \hat{u}_s} = (X_f - 1) PT_f = \alpha_f \left(\frac{u_{TB}}{1 - u_{TB}} \right) EPT_{TB} * r_{TB}|_{at \hat{u}_s}. \quad (17)$$

Therefore,

$$\begin{aligned}
\alpha_f|_{at\hat{u}_s} &= \frac{1 - u_{TB}}{u_{TB}} (X_f - 1) \frac{PT_f}{r_{TB}EPT_{TB}} \\
&= \frac{1 - u_{TB}}{u_{TB}} [(r_1 f_1 X_1 + \dots + r_m f_m X_m) - 1] \\
&\quad \times \frac{PT_f}{r_{TB}EPT_{TB}} \\
&= \frac{1 - u_{TB}}{u_{TB}} \left[\sum_{\forall i} \frac{r_i EPT_i}{r_{TB}EPT_{TB}} \alpha_i \frac{u_i}{1 - u_i} \right] \quad (18)
\end{aligned}$$

where

$$X_i = \alpha_i \left(\frac{u_i}{1 - u_i} \right) \frac{EPT_i}{PT_i} + 1$$

α_i is the variability of machine i , and u_i is the utilization of machine i . According to (18), the variability of a simple factory is a function of machine utilization, effective process times, reentry frequencies and variabilities. Note that EPT , instead of PT , is used to quantify variability for both a single machine and a simple factory. Equation (18) is an important equation, since it quantifies the relationship among factory variability and the four factors.

IV. PROPERTIES OF FACTORY VARIABILITY

Since, according to (18), factory variability is a function of machine utilization, effective process times, reentry frequencies and variabilities, it has the following properties.

A. Lower Variability for Higher Utilization of the Throughput Bottleneck

Equation (18) can be written as

$$\alpha_f = \alpha_{TB} + A \left(\frac{1}{u_{TB}} - 1 \right)$$

where

$$A = \sum_{i \neq TB} \frac{r_i EPT_i}{r_{TB}EPT_{TB}} \alpha_i \frac{u_i}{1 - u_i}.$$

For

$$\begin{aligned}
&\forall i \neq TB \quad 0 \leq u_i < 1, \quad \forall EPT_i \geq 0, \quad \forall EPT_{TB} > 0, \\
&\forall r_{TB} > 0, \quad \forall r_i > 0, \quad \forall \alpha_i \geq 0 \\
&\Rightarrow 0 \leq A < \infty.
\end{aligned}$$

For any given PT_{TB}

$$\frac{\partial \alpha_f}{\partial u_{TB}} = -\frac{A}{u_{TB}^2}$$

i.e.,

$$-\infty < \frac{\partial \alpha_f}{\partial u_{TB}} < 0, \quad \forall 0 \leq u_{TB} < 1 \quad (19)$$

Corollary 1: Variability of a simple factory is monotonically decreasing with its throughput bottleneck utilization (u_{TB}).

Due to the monotonically decreasing property, X-factor of a factory becomes large only when its utilization is very high as illustrated in Fig. 4. Compared with the prediction of a G/G/1 queueing model, actual X-factor tends to be lower when factory utilization is high. This effect makes a factory with lower X-factor (at least, lower than the prediction of G/G/1) when it is operated at high utilization.

Corollary 1 can be intuitively deduced from the definitions of cycle time, (1), and variability, (9). Assume F1 and F2 are two identical factories and the only difference between them is their TB utilizations. If the TB utilization of F1 is twice of F2s, the TB queueing time of F1 will be also longer than that of F2, which is characterized by (2). However, the nonbottleneck queueing times of these two factories are still identical since their nonbottleneck utilizations are distributed the same. Therefore, even though the system utilization (i.e., TB utilization) of F1 is larger, the total queueing time change of F1 does not change proportionally. Based on (9), the system variability must become smaller.

B. Higher Variability for Higher Utilization of the Nonthroughput Bottleneck

For a simple factory with m machines, and for the i th machine, $i = 1, 2, \dots, k-1, k+1, \dots, m$, ($i \neq k$), and k is not the throughput bottleneck machine ($k \neq TB$). Then

$$\begin{aligned} \alpha_f &= \frac{1 - u_{TB}}{u_{TB}} \left[\sum_{\forall i, i \neq k} \frac{r_i EPT_i}{r_{TB} EPT_{TB}} \alpha_i \frac{u_i}{1 - u_i} \right. \\ &\quad \left. + \frac{r_k EPT_k}{r_{TB} EPT_{TB}} \alpha_k \frac{u_k}{1 - u_k} \right] \\ \frac{\partial \alpha_f}{\partial u_k} &= \frac{1 - u_{TB}}{u_{TB}} \left(\frac{r_k EPT_k}{r_{TB} EPT_{TB}} \right) \alpha_k (1 - u_k)^{-2}. \end{aligned}$$

For

$$\begin{aligned} 0 \leq u_{TB} < 1, \quad \forall EPT_k > 0, \quad \forall EPT_{TB} > 0, \quad \forall \alpha_k > 0, \\ \forall r_{TB} > 0, \quad \forall r_k > 0 \quad 0 \leq u_k < 1 \\ \frac{\partial \alpha_f}{\partial u_k} &> 0. \end{aligned} \quad (20)$$

Corollary 2: Variability of a simple factory is monotonically increasing with its nonbottleneck utilization.

Due to the monotonically increasing property, for two factories which have the same bottleneck utilization, the one with the greater nonbottleneck capacity has the lower variability.

C. Gap Effect

Assume F1 and F2 are two identical factories and the only difference between them is the utilization of the TB machine b or the utilization of the nonbottleneck machine k . Except for b and k , utilizations of all other machines are identical. The utilizations of machine b and k in F1 are u_{b1} and u_{k1} and those of F2 are u_{b2} and u_{k2} . The utilization gap is defined as “ $u_b - u_k$.”

According to corollary 1 and 2, when u_{b2} and u_{k2} satisfy one of the following three conditions, F2 has the lower variability:

- 1) $u_{k2} < u_{k1}$ and $u_{b2} = u_{b1}$;
- 2) $u_{k2} = u_{k1}$ and $u_{b2} > u_{b1}$;
- 3) $u_{k2} < u_{k1}$ and $u_{b2} > u_{b1}$.

Corollary 3: For two identical simple factories, when one utilization gap inwardly bounds the other, the one with larger utilization gap has the lower variability.

Based on Corollary 3, increasing more nonbottleneck capacity leads to a lower α_f . Therefore, under the goal of same cycle time, factory utilization is allowed to be increased, which leads to a higher throughput rate as illustrated in Fig. 5. This conclusion is different from the Theory of Constraint (TOC), since TOC declares that only increasing the bottleneck capacity can achieve a higher system throughput rate [28].

However, increasing bottleneck capacity leads to both lower factory utilization and higher α_f . Lower utilization leads to a shorter cycle time, while higher α_f leads to a longer cycle time. When the increase in α_f is large and the cost of increasing bottleneck capacity is high, it may no longer be a wise decision to increase the throughput rate by raising bottleneck capacity.

When the factory throughput rate has to be increased, either the bottleneck or nonbottleneck machines are potential candidates to be considered for capacity expansion. In order to make the correct decision regarding capacity expansion plans, utilization changes, variability changes, and machine costs must all be taken into account.

D. Upper Bound of Variability

According to Corollary 3, factory variability is maximized when the utilization gap is zero. Assume all machine utilization is equal in a simple factory, i.e., $u_1 = u_2 = \dots = u_m = u$. Since in a simple factory system, when machine utilization is equal, based on (3), $r_1 * EPT_1 = r_2 * EPT_2 = \dots = r_m * EPT_m$. Therefore, (18) can be re-expressed as

$$\begin{aligned}
\alpha_f &= \frac{1 - u_{TB}}{u_{TB}} \\
&\times \left[\left(\frac{r_1 EPT_1 \alpha_1 + \dots + r_m EPT_m \alpha_m}{r_{TB} EPT_{TB}} \right) \left(\frac{u}{1 - u} \right) \right] \\
&= \frac{r_1 EPT_1 \alpha_1 + \dots + r_m EPT_m \alpha_m}{r_{TB} EPT_{TB}} \\
&= \sum_{i=1}^m \alpha_i, \tag{21}
\end{aligned}$$

Corollary 4: Variability of a simple factory is upper-bounded by $\sum \alpha_i$.

E. Lower Bound of Variability

According to Corollary 3, factory variability is minimized when the utilization gap is one, i.e., $u_{TB} = 1$ and $u_{NTB} = 0$, where $NTB = 1, 2, \dots, m$, except for TB .

Since

$$\alpha_f = \alpha_{TB} + A \left(\frac{1}{u_{TB}} - 1 \right)$$

where

$$A = \sum_{i \neq TB} \frac{r_i EPT_i}{r_{TB} EPT_{TB}} \alpha_i \frac{u_i}{1 - u_i}.$$

For

$$0 \leq A < \infty \quad \text{and} \quad 0 \leq u_{TB} < 1$$

as

$$\begin{aligned}
&\min \left\{ A \left(\frac{1}{u_{TB}} - 1 \right) \right\} = 0 \\
&\Rightarrow \alpha_f|_{\min} = \alpha_{TB}. \tag{22}
\end{aligned}$$

Corollary 5: Variability of a simple factory is lower-bounded by its throughput bottleneck variability (α_{TB}).

V. CONCLUSION

In this paper, variability of a simple factory is examined to establish an expression for its computation. The obtained result is then analyzed to reveal five inherent properties that describe the basic characteristics of variability for a factory. Rather than giving a quantitative prediction of the dynamics of factory behaviors under all circumstances, the descriptions revealed by these basic properties give general directions to the manufacturing line managers on how variability affects their production performances and thus how to steer their efforts. Because the method of analysis employed in this study

is through a series of rigorous steps of logical arguments, some of the results may appear to contradict established experiences on first glimpse. The impact is, however, expected to generate new discussions in a field in which there exists a huge gap between practical experiences (intuitive or otherwise) and academic researches (realistic or laboratory-wise). It is hoped that, through further coherent investigative efforts, the currently derived results can be applied to many subjects in manufacturing management, such as factory performance analysis, capacity planning, and cycle time reduction.

Although the derivation of factory variability in this paper is simply based on the G/G/1 queue, studies can be extended to the G/G/m queue [15]:

$$QT(G/G/m) = \left(\frac{C_a^2 + C_e^2}{2} \right) \left(\frac{u(\sqrt{2(m+1)}-1)}{m(1-u)} \right) EPT \quad (23)$$

where m is the number of servers. Based on (23), (18) has to be modified. The modification will reveal that the larger the number of servers is, the lower the system variability will be. The first three corollaries will still hold if combined with certain conditions; furthermore, the measure of bounds has to be modified accordingly. Detailed examination of a multiple server system is left for future research.

It may be somewhat surprising to find that variability of a factory changes with product mixes and machine utilization, rather than being independent of these factors. Declaring the factory cycle time is shorter and/or utilization is higher (i.e., lower variability) under a specific condition may not guarantee that a specific production method outperforms the others under all conditions.

Although applying queueing model to assess the performance of complex semiconductor fabs on system levels is well recognized in the semiconductor industry, on real shop floors, the actual dispatching decisions of operators are likely to blur the prediction capability of any queueing models if such actions were not considered thoroughly. Under the influences of the dispatching logic, flow variability may no longer be independent of process time variability when the operator is primarily concerned with maintaining a smooth production line. Queued lots would be sent to one of its downstream machines when WIP level at that downstream machine is low or sent to another when its WIP level is high. Furthermore, a higher WIP level may lead to a higher or lower flow variability depending on the effectiveness of the dispatching logic, especially when the behaviors of batch-forming or setup are involved.

Many heuristic sequencing rules and work release policies have been developed for semiconductor manufacturing [26], [27]. However, on shop floors, the actual dispatching decisions by the operators are usually more intricate and dynamic. Each of the developed rules may only perform well under certain (laboratory) conditions. Apart from the intuitive comparisons of the ad hoc schemes, none of them ever demonstrated the existence of a global optimal solution nor did they guarantee convergence toward such optimality [33] in their situations. The practical conditions are so dynamic that the scheduling issue is now still classified as a NP-hard problem. Before the optimal and

general solution is discovered or, at least, shown to be approachable, any attempt to develop a general queueing model to describe the correct behaviors for all semiconductor wafer fabs is very likely not amenable. Verification of the results therefore calls for the conduction of simulation studies for each specific fab under certain conditions. This is a topic to be addressed in a later publication.

ACKNOWLEDGMENT

The author would like to thank Dr. K. Hui of TSMC for his helpful guidance. This paper would have had no chance of being published without his support. He also helped to derive the mathematical expression defining relationship between cycle time and WIP. The author is also grateful to W.-Y. Chen, Director of MITD in TSMC, for his encouragement in this endeavor. The author would like to thank the Associate Editor and referee for their comments. Their suggestions of effective process time lead to the clarifications of definitions employed for utilization and X-factor.

REFERENCES

- [1] (2002, Jul.) Equipment Engineering Capabilities Guidelines (Phase 2.5). International SEMATECH. [Online]<http://www.selete.co.jp/SeleteHPJ1/Data/200207/0207b01.pdf>
- [2] L. Sattler, "Using queueing curve approximations in a fab to determine productivity improvements," in *Proc. IEEE/SEMI ASMC*, Nov. 1996, pp. 140–145.
- [3] F. G. Boebel, "Cycle time reduction program at ACL," in *Proc. IEEE/SEMI ASMC*, Nov. 1996, pp. 12–14.
- [4] J. W. Fowler, S. Brown, H. Gold, and A. Schoeming, "Measurable improvements in cycle-time-constrained capacity," in *Proc. IEEE ISSM*, Oct. 1997, pp. A21–A24.
- [5] D. W. Collins, K. Williams, and F. C. Hoppensteadt, "Implementation of minimum inventory variability scheduling 1-step ahead policy in a large semiconductor manufacturing facility," in *Proc. 6th ETFA Int'l Conf.*, Sep. 1997, pp. 497–504.
- [6] L. Sattler and R. Schlueter, "Semiconductor metrics: conflicting goals or increasing opportunities," in *Proc. IEEE/SEMI ASMC*, Sep. 1998, pp. 23–25.
- [7] O. Ruelle, "Continuous flow manufacturing: the ultimate theory of constraints," in *Proc. IEEE/SEMI ASMC*, Sep. 1997, pp. 216–221.
- [8] G. W. Horn and W. A. Podgorski, "A focus on cycle time-vs.-tool utilization "paradox" with material handling methodology," in *Proc. IEEE/SEMI ASMC*, Sep. 1998, pp. 405–412.
- [9] P. Marcoux, M. McClintock, D. Martin, and R. Woods, "Determining capacity loss from operational and technical deployment practices in a semiconductor manufacturing line," in *Proc. IEEE ISSM*, 1999, pp. 3–5.
- [10] O. Rose, "The Shortest Processing Time First (SPTF) dispatch rule and some variants in semiconductor manufacturing," in *Proc. Winter Sim. Conf.*, vol. 2, 2001, pp. 1220–1224.
- [11] M. Kishimoto, K. Ozawa, K. Watanabe, and D. Martin, "Optimized operations by extended x-factor theory including unit hours concept," *IEEE Trans. Semicond. Manuf.*, vol. 14, no. 3, pp. 187–195, Aug. 2001.
- [12] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [13] S. M. Ross, *Introduction to Probability Models*, 7th ed. Boston, MA: Academic, 2000.
- [14] J. A. Buzacott and J. G. Shanthikumar, *Stochastic Models of Manufacturing Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
- [15] W. J. Hopp and M. L. Spearman, *Factory Physics*. Chicago, IL: Irwin, 1996.

- [16] D. P. Martin, "Key factors in designing a manufacturing line to maximize tool utilization and minimize turnaround time," in *Proc. IEEE/SEMI Int. Semicond. Manuf. Sci. Symp.*, Jul. 1993, pp. 48–53.
- [17] —, "The advantages of using short cycle time management instead of continuous flow management," in *Proc. IEEE/SEMI ASMC*, Sep. 1998, pp. 43–49.
- [18] K. Ozawa, H. Wada, and T. Yamaguchi, "Optimum tool planning using the X-factor theory," in *Proc. IEEE ISSM*, 1999, pp. 49–52.
- [19] A. K. Schoemig, "On the corrupting influence of variability in semiconductor manufacturing," in *Proc. Winter Sim. Conf.*, vol. 1, 1999, pp. 837–842.
- [20] J. Robinson, "FabTime characteristic curve generator," *FabTime Cycle Time Management Newsletter*, vol. 2, no. 7, Aug. 2001.
- [21] P. Sungmin, G. T. Mackulak, and J. W. Fowler, "An overall framework for generating simulation-based cycle time-throughput curves," in *Proc. Winter Sim. Conf.*, vol. 2, 2001, pp. 1178–1187.
- [22] D. Nazzal and M. Mollaghasemi, "Critical tools identification and characteristics curves construction in a wafer fabrication facility," in *Proc. Winter Sim. Conf.*, vol. 2, 2001, pp. 1194–1199.
- [23] F. Boebel and S. Halmel, "The SIEMENS productivity program methodology and advances," presented at the ASMC, 1998, Session 6—International Session.
- [24] D. P. Connors, G. E. Feigin, and D. D. Yao, "A queueing network model for semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 9, no. 3, pp. 412–427, Aug. 1996.
- [25] S. Kumar and P. R. Kumar, "Queueing network models in the design and analysis of semiconductor wafer fabs," *IEEE Trans. Robot. Autom.*, vol. 17, no. 5, pp. 548–561, Oct. 2001.
- [26] L. M. Wein, "Scheduling semiconductor wafer fabrication," *IEEE Trans. Semicond. Manuf.*, vol. 1, no. 3, pp. 115–129, Aug. 1988.
- [27] C. Lozinski and C. R. Glassey, "Bottleneck starvation indicators for shop floor control," *IEEE Trans. Semicond. Manuf.*, vol. 1, no. 4, pp. 147–153, Nov. 1988.
- [28] E. M. Goldratt and J. Cox, *The Goal*. Great Barrington, MA: North River, 1992.
- [29] D. P. Martin, "How the law of unanticipated consequences can nullify the theory of constraints: the case balanced capacity in a semiconductor manufacturing line," in *Proc. IEEE/SEMI ASMC*, Sep. 1997, pp. 380–385.
- [30] J. H. Jacobs, L. F. P. Etman, E. J. J. van Campen, and J. E. Rooda, "Characterization of operational time variability using effective process time," *IEEE Trans. Semicond. Manuf.*, vol. 16, no. 3, pp. 511–520, Aug. 2003.

- [31] *Book of SEMI Standards (SEMI E10-0701)*. Mountain View, CA: SEMI, 2001, pp. 1–22.
- [32] J. D. C. Little, “A proof of the queueing formula: $L = \lambda W$,” *Operations Research*, vol. 9, pp. 383–387, 1961.
- [33] M. Mittler and A. K. Schoemig, “Comparison of dispatching rules for semiconductor manufacturing using large facility models,” in *Proc. Winter Sim. Conf.*, vol. 1, 1999, pp. 709–713.

List of Figures

Fig. 1. Characteristic curve showing queueing time versus machine utilization. Longer queueing time is always the price of higher utilization.

Fig. 2. Relationship between cycle time and WIP level.

Fig.3. Simple factory system versus single machine system at a specific utilization distribution. On the left is a simple factory system, and the right is a single machine.

Fig. 4. Illustration of Corollary 1: factory X-factor and variability as a function of the TB utilization. The higher u_{TB} is, the lower the α_f is.

Fig. 5. Impact of variability reduction on factory utilization for the goal of fixed cycle time.

Fig. 1

Fig. 2

Fig. 3

Fig. 4

Fig. 5

Kan Wu was born in Taipei, Taiwan, R.O.C., in 1970. He received the B.S. degree in nuclear engineering from National Tsing Hua University, Hsinchu, Taiwan, R.O.C., in 1994, the M.S. degree in industrial engineering and operations research and the M.E. degree in nuclear engineering from the University of California, Berkeley, in 1996.

He was an engineer with Tefen, Ltd., from 1996 to 1998, and with Taiwan Semiconductor Manufacturing Company from 1998 to 2003. Currently, he is an IE manager at Inotera Memories Inc., Taoyuan, Taiwan, R.O.C. His research interests include production planning, scheduling, and dispatching in semiconductor industry.