Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds

THE EFFECTIVENESS OF VARIABILITY REDUCTION IN DECREASING WAFER FABRICATION CYCLE TIME

Israel Tirkel

Department of Industrial Engineering and Management
Ben-Gurion University of the Negev
Beer-Sheva, 84105, ISRAEL

ABSTRACT

Fab operations management strives to decrease cycle-time (CT) for driving low inventory, improved quality, short time-to-market and lower cost. This work studies factors contributing to production variability, and evaluates the variability's influence on CT. It relies on queueing networks, CT and variability approximations, operational curve modeling, and common practice. It demonstrates that increasing variability drives longer CT at a growing pace, and has a larger effect on CT than utilization. Growing machine inventory weakens the impact of utilization on CT and almost eliminates it at high inventory, while the impact of variability on CT remains significant. Decline of machine availability prolongs CT at a growing pace, and is affected by variability more than utilization. Overall the primary factor of production variability is attributed to machine availability, and specifically to repair time. Reducing variability for achieving decreased CT is less costly and more effective than reducing machine utilization or increasing capacity.

1 INTRODUCTION

1.1 Motivation and goal

Fab operations management aims to decrease the production line cycle time (CT) given the required throughput. Reduced CT drives low inventory, improved quality, short time-to-market and lower cost. The production steps designated for CT improvement are usually the steps with the highest load (i.e. traffic intensity), referred to as bottlenecks. Common strategies for reducing the CT of bottlenecks call for decreasing the load, via service time improvement or machine capacity increase. Other CT improvement activities which focus on less significant factors (e.g. transfer time) are not discussed here. This work suggests to redirect the CT reduction strategies to decreasing the production step's variability rather than decreasing the load, referred to as utilization. It illustrates that given high variability, which characterizes wafer fabrication systems, decreasing the variability is more effective than decreasing the machine utilization. The study relies on queueing networks, CT and variability approximations, operational curve modeling, and common practice. Although the results are based on analytical approximations and not precise models, the demonstrated trends comprise a good basis for the analysis and the redirection of the operations management strategies.

1.2 Literature

Queuing has been the basis for many studies of production systems in general, and in semiconductor manufacturing specifically. CT is one of the top performance measures investigated in literature and practice, due to its impact on operations and business. A common method for demonstrating the trends of CT in a production system is the operating curve. It exhibits the tradeoff between CT and utilization or

throughput, based on queueing (Aurand and Miller 1997, Veeger *et al.* 2010). The simplest illustration of the operating curve relies on an M/M/1 queueing model which shows that CT increases with utilization at a growing pace. It demonstrates the function $CT = 1/(\mu - \lambda)$, where CT is the mean CT, λ is the mean arrival rate and μ is the mean service rate.

The M/M/1 fundamental model has been extended to various G/G/m models, considering general arrival and service distributions and m servers, where $m \geq 1$ and an integer. The extended models approximate the mean CT using the coefficient of variation (CV), defined by the ratio between the standard deviation and the mean, of the inter-arrival time and the service time. The CV substitutes the use of specific distributions by representing their variability. Further generalized and more complex models consider, in addition, partial machine availability (i.e. less than 100%) and approximate the mean CT using expressions for availability (A), repair time duration, and repair time CV.

The first study of a G/G/1 model using CV's of inter-arrival and service times was presented by Kingman (1961). It was followed by the G/G/1 studies of Shanthikumar and Buzacott (1980), and Whitt (1983). Extended approximation to more than a single server G/G/m model was developed by Sakasegawa (1977), and followed by Whitt (1993), and Buzacott and Shanthikumar (1993). Hopp and Spearman (2001) presented comprehensive G/G/m approximations, including a model with partial machine availability. Based on their model Morrison and Martin (2007) investigated various cases of specific processing types (e.g. parallelism, idle with work). See Table 1 for the models comparison.

Table 1: CT approximation models.

A	m	Kingman (1961)	Whitt (1993)	Shanthikumar and Buzacott (1993)	Hopp and Spearman (2001)	Morrison and Martin (2007)
100%	1	√	√	√	√	√
	>1		√	√	√	√
<100%	1				√	√
	>1				√	√

Wafer fabrication is a complex manufacturing system subject to high variability which significantly affects CT. The variability in the production system is generated by numerous components considered in this work. The analysis presented relies on G/G/m queueing models with repair time, based on Hopp and Spearman (2001) and supported by Morrison and Martin (2007). The generalized CT approximations include inter-arrival time variability, service time variability, more than a single machine, partial availability, and repair time variability. The work challenges the claim that "utilization has a more dramatic effect on CT than variability" (Hopp and Spearman 2001). It demonstrates the significant effect of variability on CT, and indicates it can exceed the effect of utilization. It also explains how reducing CT by decreasing variability is more effective than by decreasing utilization.

The rest of this work is organized as follows. The production model is described in Section 2, the variability and CT queueing approximations are presented in Section 3, the CT reduction results and analysis using operating curves is explained in Section 4, and the summary and concluding remarks in Section 5.

2 PRODUCTION MODEL

2.1 Production system

The wafer fabrication line modeled in this work is based on a queueing network in tandem. Figure 1 illustrates a production system with N steps,

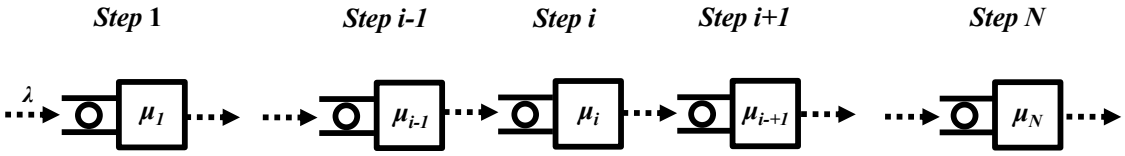

Figure 1: Production system model.

where,
λ is the mean arrival rate to the system [items/time-unit],
μ_i is the mean service rate in *Step i* [items/time-unit],

2.2 Machine utilization

Machine utilization (u) is the proportion of time a machine is busy (i.e. processing). Literature and practice use two definitions for utilization. The first, considers the proportion of the time busy and the total time, such that $u = \lambda/\mu$, and $u = \lambda/(\mu m)$ for more than a single machine. This definition is adequate, for example, in economic models which consider the utilization of a machine capital cost. The second, considers the proportion of time busy and the time the machine is available for production, such that $u = \lambda/(\mu A m)$. This definition is adequate, for example, in production models which consider the traffic intensity of a machine. Clearly, when $A = 100\%$ both definitions are identical. This work applies the latter definition.

2.3 Machine availability

Machine availability is the proportion of time a machine is available for running production. It is defined by $A = f/(f + r)$, where f is the mean time to failure (MTTF) or the operating time, and r is the mean time to repair (MTTR) or the repair time. The availability increases with prolonged operating time or reduced repair time. The variability of the availability depends on the durations and variability of each, the MTTF and the MTTR. Lower availability generates higher variability in the production system. Higher variability, of the operating time or the repair time, consequently also increases the system's variability and prolongs the mean CT. This work considers the variability generated by the machine availability due to two factors: (1) the availability – change of the repair time duration given the operating time; longer repair time will reduce the availability and generate higher variability, and (2) the repair time variability – change of the repair time variability given the repair time duration; higher repair time variability will generate higher variability in the production system.

2.4 Cycle Time

The mean CT of k items (indexed j=1, 2…, k) in any single *Step i* is defined by,

$$CT_i = \frac{1}{k}\sum_{j=1}^{k} CT_{ij} = \frac{1}{k}\sum_{j=1}^{k}\left(CompleteTime_{i,j} - StartTime_{i,j}\right) \quad (1)$$

where,
$StartTime_{1,j} = 0$ and $StartTime_{i,j} = CompleteTime_{i-1,j}$.

The mean CT of N steps in a production line is defined as the sum of the individual steps mean CT, such that $CT = \sum_{i=1}^{N} CT_i$. The major CT components in any step include waiting-in-line time, service time and additional times (e.g. transfer time, set-up time, waiting for batch time). The transfer time depends on the material handling system, it is usually short and disregarded here. The set-up time is assumed to be included in the service time. The waiting for batch time is disregarded due to its low frequency in the production line. Figure 2 illustrates the significant CT components considered in this work.

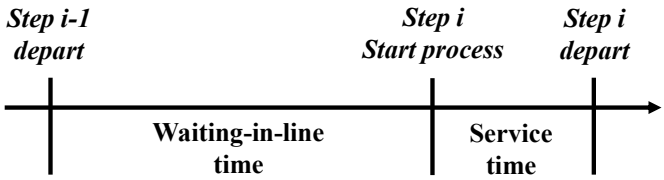

Figure 2: CT major components breakdown.

The waiting-in-line time is an unproductive time that extends CT. Longer CT prolongs time-to-market, generates excess work-in-process, drives lower quality and incurs extra costs. The X-Factor is a measure of performance [scalar] indicating the relative extent of the CT. It is defined as the ratio between the CT and the service time, where X-Factor ≥ 1. Higher X-Factor indicates poorer performance. In the operating curves illustrated in Section 4, the service time is 1 time-unit and the mean CT demonstrated also reflects the X-Factor.

3 VARIABILITY AND CT APPROXIMATIONS

3.1 Variability

Due to the complexity of semiconductor manufacturing, there are numerous factors impacting the variability of the production system, such as: machine performance, process technology, delivery schedule, material handling and logistic systems. This study considers the variability of the inter-arrival time, service time, inter-departure time, availability and repair time. The various CV's are defined as follows:

- $\mathrm{Ca_i}$ is the CV of the inter-arrival time of Step i,
- $\mathrm{Cs_i}$ is the CV of the service time of Step i,
- $\mathrm{Cd_i}$ is the CV of the departure time of Step i, and
- $\mathrm{Cr_i}$ is the CV of the repair time of Step i.

The CV of the inter-arrival time of the first step is very low ($Ca_1 \to 0$), since it is externally controlled. The inter-departure time CV of *Step i*, is approximated (Whitt 1983) in the case of a single server by,

$$Cd_i{}^2 = u_i{}^2 Cs_i{}^2 + (1 - u_i{}^2) Ca_i{}^2 \tag{2}$$

At very high utilization ($u \to 1$), Cd_i converges to Cs_i. At very low utilization ($u \to 0$), Cd_i converges to Ca_i. Overall, the upper bound of Cs_i in a tandem queuing network is the $Maximum\{Cs_i\}, \forall\ i = 1, 2, \ldots N$. Consequently, the inter-departure times and the inter-arrival times CV's are similarly bounded. In a case where all Cs_i are equal, all Ca_i will be equal as well (except at the first steps where is Ca_i is lower), such that $Ca_i = Cs_i, \forall\ i = 2,3, \ldots N$, where $u \to 1$. For practicality, the illustrations in Section 4 assumes the inter-arrival and service times CV's are equal. The case of numerous servers ($m \geq 1$) is approximated by,

$$Cd_i{}^2 = 1 + (1 - u_i{}^2)\left(Ca_i{}^2 - 1\right) + \frac{u_i{}^2}{\sqrt{m_i}}\left(Cs_i{}^2 - 1\right) \tag{3}$$

Here, the upper bound of Cs_i is identical to the single server case. The inter-departure and the inter-arrival times CV's are upper bounded by $\left(1 + \left(Cs_i{}^2 - 1\right)/\sqrt{m_i}\right)^{1/2}$ which less than Cs_i. Thus, the above upper bounds still hold for the general case ($m_i = 1,2, \ldots$).

Hopp and Spearman (2001) categories the levels of variability as follows:
Low variability – CV less than 0.75,

Medium variability – CV between 0.75 and 1.33, and
High variability – CV more than 1.33.
Various studies (Rose *et al.* 2000, Jacobs *et al.* 2003, Akhavan-Tabatabaei *et al.* 2009) consider wafer fabrication variability to range from 0.75 through 3.74. Consequently and based on our experience, the analysis in Section 4 considers CV's range from 0.5 through 3.5.

3.2 Queueing approximations

Based on Kingman (1961), Sakasegawa (1977), and Whitt (1983), the mean CT approximation in a G/G/*m* queueing system, excluding the effect of availability, is expressed as follows,

$$CT = \frac{1}{\mu} + \left(\frac{C_a^2 + C_s^2}{2}\right)\left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right)\frac{1}{\mu} \tag{4}$$

Hopp and Spearman (2001) considered the effect of partial machine availability in their CT approximation by defining the effective service time as follows by, $t_e = 1/\mu A$. Consequently, the effective mean service rate can be expressed by $\mu_e = m\mu A$. The CV of the effective service time is expressed as follows,

$$C_e^2 = C_s^2 + (1 + C_r^2)A(1 - A)r\mu \tag{5}$$

Based on equation (4) and Hopp and Spearman's (2001), the mean CT approximation in a G/G/*m* queueing system with partial availability is expressed as follows,

$$CT = \frac{1}{\mu A} + \left(\frac{C_a^2 + C_e^2}{2}\right)\left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right)\frac{1}{\mu A} \tag{6}$$

Morrison and Martin (2007) approximations exhibit similar trends to Hopp and Spearman (2001). The latter is used as the basis for the results and analysis in the rest of this work.

4 RESULTS AND ANALYSIS

4.1 CT reduction approach

The two major approaches considered for CT reduction are:

1. Reducing utilization by either (a) decreasing throughput, (b) increasing capacity, or (c) increasing availability; in all options, the working point will stay on the same operating curve (e.g. sloped dotted arrow in Figure 3).
2. Reducing variability of either (a) availability, (b) repair time, or (c) service time; in all options, the working point will move downward from one operating curve to the next (e.g. vertical dotted in Figure 3).

Hopp and Spearman (2001) claim that the two contributors to CT are utilization and variability, of which utilization has the most dramatic effect. This work shows that in the semiconductor industry variability has frequently a greater effect on CT than utilization. Moreover, it is usually more effective to decrease variability for reducing CT. The purpose of this work is to study the impact of variability and utilization reduction on mean CT. Higher utilization drives prolonged CT at a growing pace, given constant variability. Also, higher variability drives prolonged CT at a growing pace, given constant utilization. Reducing the variability will enable CT reduction or utilization increase, given fixed capacity (μ). The motivation for this work stems from the need to study the impact of variability on CT reduction, and to com-

pare the impact of variability versus the impact of utilization on CT reduction. The cost of CT reduction associated with decreasing utilization is usually very high. The cost of CT reduction associated with decreasing variability is frequently significantly lower and thus more effective.

4.2 Utilization versus variability effect on CT

Figure 3 illustrates operating curves of mean CT versus utilization based on equation (6), for G/G/1 at various variability levels and 100% availability. It exhibits that CT increases with utilization at a growing pace. Also that CT increases with variability at a growing pace, demonstrated by the increasing distance between consecutive curves. It is shown that variability has a larger effect on CT than utilization, using the following example (dotted arrows): Mean CT decrease from 30.0 to 13.0 is enabled by utilization reduction from 70% to 32% (by 54%) or CV reduction from 3.0 to 2.0 (by only 33%). Furthermore, reducing CT to less than 9.0 is feasible by only decreasing variability but isn't by only decreasing utilization. Higher effect of utilization on CT is reflected only where utilization exceeds 90%, at very low variability or at very high unrealistic CT, which do not reflect semiconductor manufacturing environment.

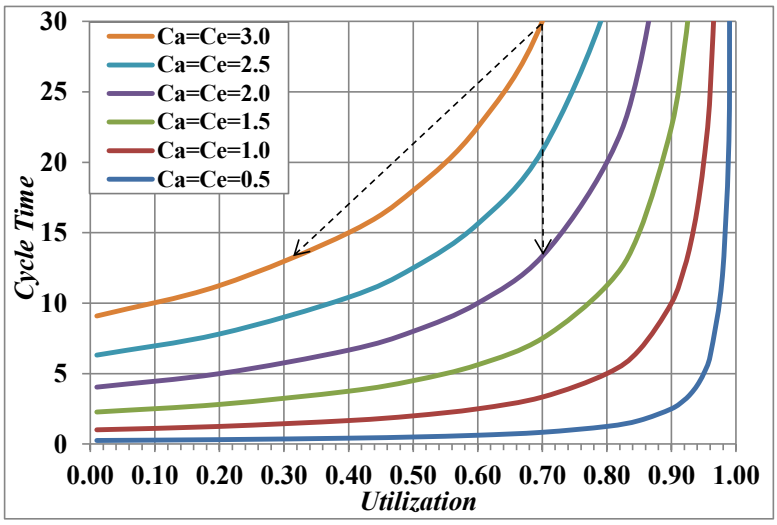

Figure 3: Utilization and variability impact on CT in G/G/1 with 100% availability.

Figure 4 illustrates operating curves of mean CT versus utilization based on equation (6), for G/G/10 at various variability levels and 100% availability. It exhibits that through 80% utilization, the curves are almost flat reflecting no effect of utilization on CT but only the effect of variability. Where utilization is above 80%, the trends are similar to Figure 3. Clearly, higher m extends the flatness of the curves.

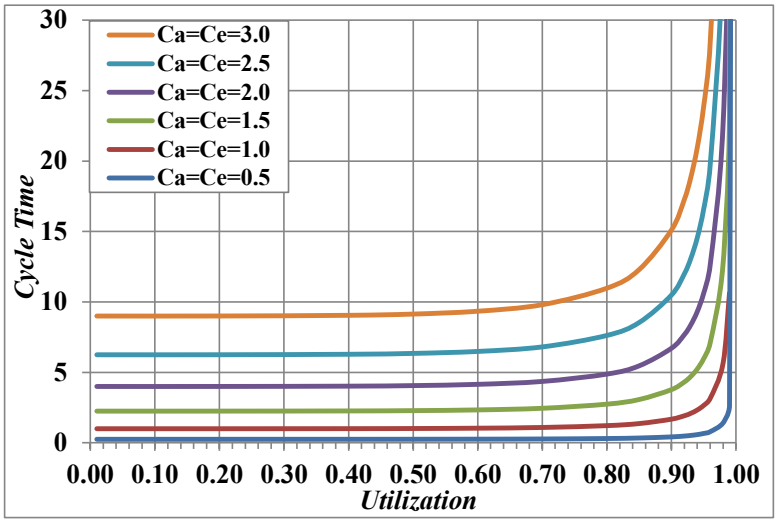

Figure 4: Utilization and variability impact on CT in G/G/10 with 100% availability.

Figure 5 illustrates operating curves of mean CT versus utilization based on equation (6), for G/G/1 at various variability levels and 90% availability. A is kept fixed at 90%, Cr is kept fixed at zero (no repair time variability), and Cs is reduced in order to generate identical Ce values to Figures 3 and 4. It exhibits that CT increases with utilization and with variability at a faster pace than in Figure 3. The effect on CT is illustrated using the following example (dotted arrows): CT decrease from 30.0 to 13.0 is enabled by utilization reduction from 67% to 25% (by 63%) or CV reduction from 3.0 to 2.0 (by only 33%). The effect of variability versus utilization on CT is larger here, and will grow further with lower availability (since CT curves grow higher with reduced availability).

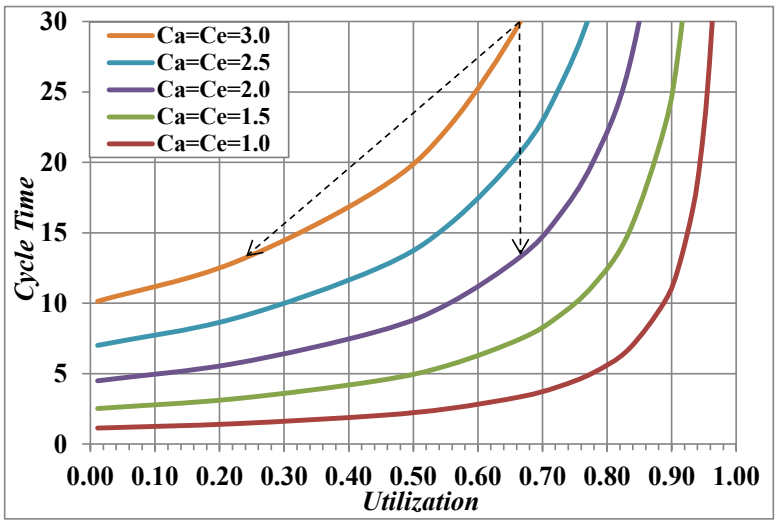

Figure 5: Utilization and variability impact on CT in G/G/1 with 90% availability.

4.3 Factors effecting variability

Figure 6 illustrates Ce based on equation (5) in G/G/1. Cs is kept at zero since its impact on Ce is linear, adding no value in this analysis. Also, the availability is kept constant at 80%. Ce grows with increasing

repair time variability. Ce also grows with increasing repair time, although the availability stays constant. As illustrated above, higher variability increases CT and reduces the relative impact of utilization on CT.

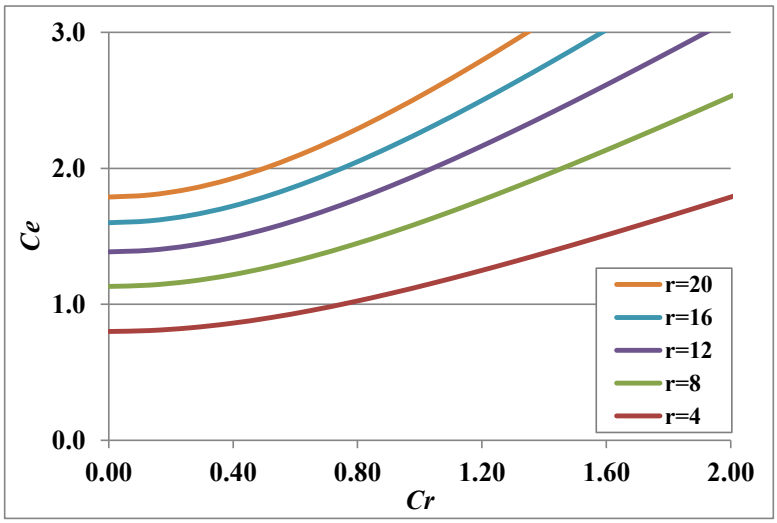

Figure 6: Ce as a function of repair time and Cr, constant availability at 80%.

The following example demonstrates the impact of decreasing variability via repair time on the mean CT. Based on Figure 6, assume that: $A = 80\%$, $u = 90\%$ and $Cr = 1.0$ (e.g. exponential). Assume the repair time is reduced from 16 to 8 time-units. As a result, the Ce is decreased from 2.26 to 1.60. Consequently, the CT is reduced from 38.25 to 22.25 which is a 42% reduction. The example exhibits that mean CT can be significantly reduced with no utilization increase. This is just by introducing shorter (by half) and more frequent (double) downtimes (i.e. separating maintenance into a few time segments).

4.4 Variability reduction

Based on Section 3.1, the variability in wafer fabrication is high and generated due to the service versus the arrival process. Based on equation (5) and Section 4.3, the sources of effective service time variability are the service time, the availability and the repair time. The service time variability is not necessarily high since the processing time is usually automated and its variability is generated due to causes such as wafer lot size (Morrison and Martin 2007). However, the impact of availability and repair time variability are significant. Lower variability can be obtained by reducing the repair time (e.g. Section 4.3 example). Also, by reducing the variability of the repair time, for example, via increasing the readiness for maintenance (e.g. labor, spares inventory). Thus, slight increase in the cost of maintenance can significantly reduce the variability of machine availability and consequently reduce the mean CT. This is versus increasing the utilization by reducing throughput or acquiring additional machine capacity which incurs higher costs. This work aims to analyze and compare the effectiveness of each CT reduction strategy, although the mechanisms for variability reduction are not always as apparent as for capacity increase.

5 SUMMARY AND CONCLUSIONS

This work presented a CT reduction approach based on operating curves and queueing approximations. The generalized mean CT models referenced included the impact of inter-arrival time variability, service time variability, machine inventory, partial availability, and repair time variability. This work challenged the claim "that utilization has a more dramatic effect on CT than variability" (Hopp and Spearman 2001) and suggested an alternate strategy for CT reduction based on decreasing variability.

It exhibited the significant effect of variability on CT, and indicated it can exceed the effect of utilization in wafer fabrication. Furthermore, it explained that decreasing CT by reducing variability can be more effective than by reducing utilization. The conclusions reflect that variability has higher effect on CT than utilization at high variability environment relevant to semiconductors manufacturing. This effect grows stronger with higher machine inventory and lower availability. The effect of utilization on CT is higher at low variability levels and at very high utilization, not reflecting wafer fabrication environment. Finally, the purpose of this work was to redirect strategies aimed to decrease CT by focusing on reducing variability, rather than reducing the production step's utilization.

REFERENCES

Akhavan-Tabatabaei R., Ding S., Shanthikumar J.G., 2009. A Method for Cycle Time Estimation of Semiconductors Manufacturing Toolsets with Correlations. *Proceedings of the Winter Simulation Conference* 1719-1729.

Aurand S.S., Miller P.J. 1997. The Operating Curve: A Method to Measure and Benchmark Manufacturing Line Productivity. Advanced Semiconductor Manufacturing Conference.

Buzacott J.A., Shanthikumar J.G. 1993. Stochastic Models of Manufacturing Systems. Prentice Hall, New-Jersey.

Hopp W.J., Spearman M.L. 2001. Factory physics. McGraw-Hill, Boston.

Jacobs J.H., Etman L.F.P., Campen E.J.J. van, Rooda J.E. 2003. Characterization of Operational Time Variability Using Effective Process Times. *IEEE Transactions on Semiconductor Manufacturing* 16(3): 352-362.

Kingman J.F.C. 1961. The Single Server Queue in Heavy Traffic. *Proceedings of the Cambridge Philosophical Society* 57: 902-904.

Morrison J.R., Martin D.P. 2007. Practical Extensions to Cycle Time Approximations for the G/G/m-queue with Applications. *IEEE Transactions on Automation Science and Engineering* 4(4): 523–532.

Rose O., Duemmler M., Schoemig A. 2010. On the Validity of Approximation Formulae for Machine Downtimes. Research Report Series, Institute of Computer Science at the University of Würzburg, Germany, Report no. 250.

Sakasegawa H. 1977. An Approximation Formula $Lq=\alpha\beta(1-p)$. *Annals of the Institute for Statistical Mathematics* 29: 67–75.

Shanthikumar J.G., Buzacott J.A. 1980. On the Approximations to the Single Server Queue. *International Journal of Production Research* 18(6): 761-773.

Veeger L.F.P., Etman J.V.H., Rooda J.E. 2010. Generating Cycle Time-Throughput Curves Using Effective Process Time Based Aggregate Modeling. *IEEE Transactions on Semiconductor Manufacturing* 23(4): 517-526.

Whitt W. 1983. The Queueing Network Analyzer. *Bell System Technology Journal* 62(9): 2779-2815.

Whitt W. 1993. Approximating the GI/G/m queue. *Production and Operations Management* 2(2): 114-161.

AUTHOR BIOGRAPHY

ISRAEL TIRKEL is a faculty researcher-lecturer in the department of Industrial Engineering and Management at Ben-Gurion University of the Negev, Israel. He has worked for Intel Corporation, in Israel and the USA, for twenty-three years in senior management positions of Fab Operations and Program Management. He received his B.Sc. with distinction at 1983, M.Sc. with distinction at 2009, and Ph.D. at 2011 in Industrial Engineering and Management, from Ben-Gurion University of the Negev. His areas of specialization are production and operations analysis and management, and project management, which he formerly practiced and is now investigating and lecturing. His prior research include work in coopera-

tion with companies such as Intel, Micron and KLA. He is an Associate Editor in IEEE Transactions on Semiconductors Manufacturing. His e-mail address is tirkel@bgu.ac.il.