Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

SINGLE-SERVER AGGREGATION OF A RE-ENTRANT FLOW LINE

Casper Veeger
Pascal Etman
Jacobus Rooda

Eindhoven University of Technology
Den Dolech 2, Whoog 0.127
5600MB Eindhoven, THE NETHERLANDS

Ivo Adan

Eindhoven University of Technology
Den Dolech 2, HG 9.07
5600MB Eindhoven, THE NETHERLANDS

ABSTRACT

On-time delivery performance of a semiconductor manufacturing system depends on the cycle time distribution of lots produced in the manufacturing network. A detailed simulation model of the manufacturing system that can predict the cycle time distribution may be helpful in performance improvement activities, but requires considerable development and maintenance effort. To reduce development and maintenance effort, an aggregate model has recently been developed that is a lumped-parameter representation of a manufacturing workstation. The lumped-parameters are directly determined from arrival and departure events measured at the workstation in operation. In this paper, we investigate under which conditions the previously developed aggregate model can be used to model a re-entrant flow line of workstations, motivated by semiconductor manufacturing. We find that the range of throughput levels for which accurate cycle time predictions are obtained decreases for increasing network size.

1 INTRODUCTION

On-time delivery performance, and time-to-market of new products are key to the profitability of many semiconductor facilities, and may be improved by reduction of the cycle time (i.e., time a lot spends in the manufacturing system). Queueing models may be used to assess the effect of operational and planning decisions on the cycle time distribution. Two types of queueing models can be distinguished: analytical models and discrete-event simulation models.

Analytical queueing network models represent the manufacturing system as a group of nodes, where each node typically represents a workstation. Shanthikumar et al. (2007) give an overview of queueing network models that can be used to model semiconductor facilities, but also state that the use of queueing theory has been considered unsatisfactory so far (Shanthikumar et al. 2007), because it is difficult to include many factory-floor details in an analytical model. Furthermore, analytical queueing models are typically used to predict the first moment (the mean) of the cycle time distribution, and not the whole cycle time distribution.

An alternative to model semiconductor manufacturing systems is discrete-event simulation modeling, which allows for the inclusion of all relevant factory-floor aspects required to accurately predict the cycle time distribution. Examples of such approaches include Miller (1990) and Kiba et al. (2009). Because many factory-floor aspects may be relevant in semiconductor factories, detailed simulation models often require much development time and maintenance, and model evaluations are typically computationally expensive.

To reduce computation time, a technique that may be used is aggregation. Aggregation combines several system components in a single component that has similar behavior. For example, Brooks and Tobias (2000), Johnson et al. (2005) used a simplification technique in which non-bottleneck workstations were replaced by a constant delay. Rose (2007) modeled the non-bottleneck workstations by a FCFS (First-Come-First-Served) single-server system with Work-In-Process (WIP)-dependent process times, which are determined by running a full-detail simulation model at various

978-1-4244-9864-2/10/$26.00 ©2010 IEEE

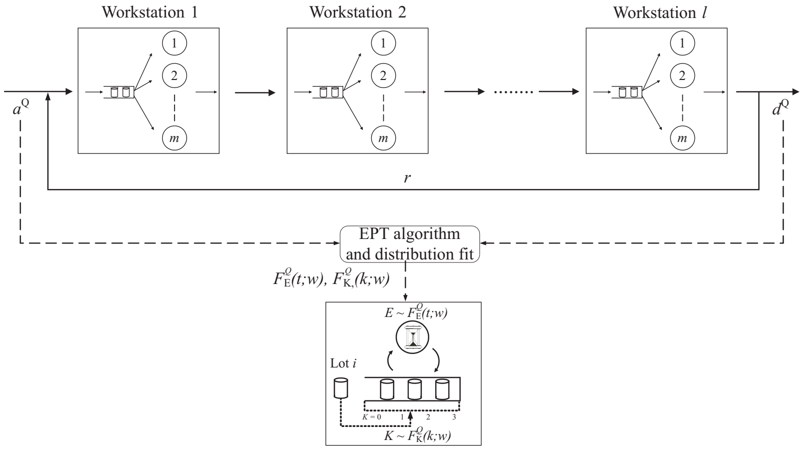


Figure 1: Modeling a manufacturing network by a single-server aggregate model

utilization levels. Rose (2007) used this aggregate model for the prediction of the cycle time distribution.

The aforementioned model abstractions require that a model of the system is available beforehand to determine the model parameters of the abstract model. In this paper, we investigate using the single server EPT-based aggregate model presented in Veeger et al. (2009a), for which the model parameters can be obtained directly from arrival and departure events measured at the manufacturing system in operation. No (detailed) simulation or analytical model of the network is needed beforehand to construct the aggregate model.

The EPT-based aggregate model developed in Veeger et al. (2009a) is a single-server representation of a manufacturing workstation with a generally distributed WIP-dependent process time distribution, which is referred to as the EPT distribution. Lots that arrive at the aggregate model have a probability to overtake one or more lots already in the system, according to a lot overtaking probability distribution. From lot arrival and departure events, we calculate the EPT realizations and the amount of overtaken lots for each lot processed at the system being modeled, which is used to estimate the EPT and overtaking distribution. The EPT-based aggregate model is tested for various workstation configurations in Veeger et al. (2009a), but not for networks of workstations.

In this paper, we investigate under which conditions the EPT-based aggregate model presented in Veeger et al. (2009a) is able to accurately predict cycle time distributions for a simulation case of a re-entrant flow line, motivated by semiconductor manufacturing. We model the entire flow line by a single-server aggregate model of the type presented in Veeger et al. (2009a). We estimate the EPT and overtaking distribution of the single server from arrival times of lots at the network, and departure times of lots from the network. Two scenarios of the flow line are investigated.

2 MODEL CONCEPT

This section describes the EPT-based aggregate modeling method developed by Veeger et al. (2009a) that we use to model the flow line. Figure 1 visualizes the aggregate modeling approach. The figure depicts the re-entrant flow line under consideration. The re-entrant flow line has l workstations that each consist of an infinite-capacity queue and m identical parallel servers. Each lot is processed r times by the flow line, after which it leaves the system.

2.1 Aggregate Model Concept

The single-server aggregate model representation of the re-entrant flow line is depicted at the bottom of Figure 1. This aggregate model is the aggregate model as presented in Veeger et al. (2009a). That is, the aggregate model consists of an infinite queue, and a timer. Lots arrive in the queue of the aggregate model according to some arrival process, as observed at the entrance of the re-entrant flow line. Lot i is defined as the i^{th} *arriving* lot at the (first buffer of) the network. Because the aggregate model represents the entire network, the queue of the aggregate model contains *all* lots that are currently in the system. Lots stay in this queue during processing. If the aggregate process time has elapsed, the lot that is currently first in the queue leaves the system. Each Lot i that arrives in the queue has a probability to overtake K other lots that are already in the system. Number of lots to overtake K is sampled from overtaking probability distribution $F_{\mathrm{K}}^{\mathrm{Q}}(k;w)$, which defines the probability $P(K \leq k;w)$ that k or fewer lots are overtaken in the network Q; probability distribution $F_{\mathrm{K}}^{\mathrm{Q}}(k;w)$ depends on the number of lots w in the queue just before Lot i arrives (so not including Lot i itself). The reason is that the amount of lots that an arriving lot can overtake depends on the number of lots already in the system.

The timer determines when the next lot leaves the queue. The timer starts when i) a lot arrives while no lots are present in the queue, or ii) a lot departs while leaving one or more lots behind. When the timer starts, a time period E is sampled from probability distribution $F_{\mathrm{E}}^{\mathrm{Q}}(t;w)$, which defines the probability $P(E \leq t;w)$ that E is less than or equal to t in network Q. The probability distribution $F_{\mathrm{E}}^{\mathrm{Q}}(t;w)$ depends on the number of lots w in the system just after the timer start. So in case of an arrival (case i)), w includes the arrived lot. In case of a lot departure (case ii)), w does not include the departed lot. Time period E is referred to as the Effective Process Time (EPT), and $F_{\mathrm{E}}^{\mathrm{Q}}(t;w)$ as the EPT distribution. When the EPT is finished, the lot that is presently first in the queue leaves the system.

The EPT is obviously not the processing time in the queueing network being modeled, but instead relates to the interdeparture times of lots from the network. The interdeparture time depends on the Work-In-Process (WIP) in the system: the more WIP in the system, the shorter the interdeparture time will typically be.

2.2 Measuring the EPT and Overtaking Distributions

The input of the aggregate model consists of EPT distribution $F_{\mathrm{E}}^{\mathrm{Q}}(t;w)$ and overtaking distribution $F_{\mathrm{K}}^{\mathrm{Q}}(k;w)$. We estimate $F_{\mathrm{E}}^{\mathrm{Q}}(t;w)$ and $F_{\mathrm{K}}^{\mathrm{Q}}(k;w)$ from measured arrival events a^{Q} at the network Q and departure events d^{Q} from the network Q, as illustrated in Figure 1. Arrival and departure events consists of the time the event occurred, and the arrival number i of the lot that arrives or departs. We measure a^{Q} and d^{Q} of a number of lots processed by the network, while it is operating at a certain throughput ratio $\delta/\delta_{\mathrm{max}}$, with δ the actual throughput, and δ_{max} the maximum obtainable throughput of the network. We refer to this throughput ratio as the training level. Restricting ourselves to measuring at a single training level reflects the situation in a real factory, where the network is also operating at a certain throughput ratio during the measurement period. From arrivals a^{Q} and departures d^{Q}, we calculate EPT realizations and overtaking realizations using the algorithm given in Appendix A. An example of the calculation of EPTs and overtaking realizations can be found in Veeger et al. (2009a).

The EPT realizations calculated by the algorithm are grouped according to the number of lots w in the system upon the EPT start. For each WIP-level w for which EPT-realizations are obtained, a distribution is estimated, which is used for the EPT distribution $F_{\mathrm{E}}^{\mathrm{Q}}(t;w)$ in the aggregate model. For the various experiments presented in this paper, we assume that the EPT distributions for each WIP-level are gamma distributed, with mean EPT $t_{\mathrm{e},w}^{\mathrm{Q}}$ and coefficient of variability $c_{\mathrm{e},w}^{\mathrm{Q}}$. In this notation, ‘e’ is a short-hand notation for EPT, w denotes the WIP level, and Q denotes the system modeled by the single-server aggregation (the network Q in this case). Overtaking realizations are also grouped, but now according to the number of lots in the system w upon arrival. For each WIP-level, we use the measured overtaking distribution directly for $F_{\mathrm{K}}^{\mathrm{Q}}(k;w)$ in the aggregate model.

The aggregate model with estimated distributions $F_{\mathrm{E}}^{\mathrm{Q}}(t;w)$ and $F_{\mathrm{K}}^{\mathrm{Q}}(k;w)$ is used to predict the mean and distribution of the cycle time of the network Q for throughput levels other than the training level.

2.3 Curve Fitting

The accuracy of the cycle time predictions by the aggregate model depends on whether EPT-distribution parameters $t_{\mathrm{e},w}^{\mathrm{Q}}$ and $c_{\mathrm{e},w}^{\mathrm{Q}}$, and overtaking distribution $F_{\mathrm{K}}^{\mathrm{Q}}(k;w)$ can be accurately estimated for the various WIP-levels. In a network, we may not be able to accurately obtain these estimates for certain WIP levels from the measured arrivals and departures, because these WIP levels were rare, or did not occur at all during the data collection period. During the data collection period, the WIP in the network will fluctuate in a range around an average WIP level that one expects for the average throughput during the collection period. Given the limited number of arrivals and departures measured in the data collection period, it is unlikely that high or low WIP levels are observed compared to the average WIP level, in particular for increasing network size. Consequently, $t^{\mathrm{Q}}_{\mathrm{e},w}$, $c^{\mathrm{Q}}_{\mathrm{e},w}$, and $F^{\mathrm{Q}}_{\mathrm{K}}(k;w)$ cannot be accurately estimated for these WIP levels.

A curve-fitting procedure is used to deal with the limited number of EPT realizations obtained in the data collection period. Closed-form expressions $\hat{t}^{\mathrm{Q}}_{\mathrm{e}}(w)$ and $\hat{c}^{\mathrm{Q}}_{\mathrm{e}}(w)$ are fitted to the measured $t^{\mathrm{Q}}_{\mathrm{e},w}$ and $c^{\mathrm{Q}}_{\mathrm{e},w}$ values. Expressions $\hat{t}^{\mathrm{Q}}_{\mathrm{e}}(w)$ and $\hat{c}^{\mathrm{Q}}_{\mathrm{e}}(w)$ are also used to estimate the mean and coefficient of variation of the EPT for WIP levels for which no $t^{\mathrm{Q}}_{\mathrm{e},w}$ and $c^{\mathrm{Q}}_{\mathrm{e},w}$ estimates have been measured. This may be viewed as extrapolation.

The expressions used for the curve fitting should be able to represent the observed functional behavior of $t^{\mathrm{Q}}_{\mathrm{e},w}$ and $c^{\mathrm{Q}}_{\mathrm{e},w}$. In the experiments performed in this paper, we typically observe that $t^{\mathrm{Q}}_{\mathrm{e},w}$ decreases for increasing w, and approaches a horizontal asymptote for $w \to \infty$. This is because for $w > 1$, the mean EPT may be interpreted as the mean interdeparture time from the network. For increasing w, more servers in the network have lots to process which results in a lower mean interdeparture time. Intuitively, we would expect that for increasing w the mean EPT approaches a minimum mean interdeparture time that corresponds to the maximum capacity of the system. We refer to this minimum interdeparture time as the expected horizontal asymptote, which is given by $1/\delta_{\max}$. Similarly, we observe from the experiments performed in this paper that $c^{\mathrm{Q}}_{\mathrm{e},w}$ approaches a horizontal asymptote for $w \to \infty$; for $w > 1$, $c^{\mathrm{Q}}_{\mathrm{e},w}$ can be interpreted as the coefficient of variation of the interdeparture time.

In the simulation experiments presented in the subsequent sections, the following reciprocal function is used for the curve fitting of $t^{\mathrm{Q}}_{\mathrm{e},w}$:

$$\hat{t}^{\mathrm{Q}}_{\mathrm{e}}(w) = \theta + \frac{\eta}{w^{\lambda}} \tag{1}$$

In this equation, θ represents the value of $\hat{t}^{\mathrm{Q}}_{\mathrm{e}}(w)$ at $w = \infty$; $\theta + \eta$ represents the value of $\hat{t}^{\mathrm{Q}}_{\mathrm{e}}(w)$ at $w = 1$. With $\eta > 0$, the curve is decreasing for increasing w and approaches a horizontal asymptote at θ for $w \to \infty$. Variable λ determines the gradient of the curve. Variables θ, η, and λ are estimated using a weighted least-squares fitting procedure: $t^{\mathrm{Q}}_{\mathrm{e},w}$ is weighted according the $\sqrt{n_w}$, which is the number of measured EPT realizations that started with WIP level w. In the curve fitting procedure, we use a lower bound of 0 for θ, η, and λ. For η, we use an upper bound equal to the sum of the mean processing times encountered by a lot processed by the flow line, which equals $l \cdot r \cdot t_0$. For θ and λ, we do not use an upper bound.

A similar expression and curve fit procedure is used for $c^{\mathrm{Q}}_{\mathrm{e},w}$. In the simulation experiments, $c^{\mathrm{Q}}_{\mathrm{e},w}$ is generally found to be either increasing or decreasing for increasing w, which gives $\eta < 0$ or $\eta > 0$ respectively. The reciprocal function (1) is the simplest function that we have found so far that is able to model the functional behavior of $t^{\mathrm{Q}}_{\mathrm{e},w}$ and $c^{\mathrm{Q}}_{\mathrm{e},w}$ reasonably well. The exponential function proposed in Veeger et al. (2009a) does not appear to be suitable in the context of flow lines.

The measured distribution $F^{\mathrm{Q}}_{\mathrm{K}}(k;w)$ is directly used in the aggregate model, without a curve-fitting procedure. For WIP levels lower than the lowest WIP level for which we measured overtaking realizations, we assume that no overtaking occurs. For WIP levels higher than the highest WIP level for which we measured overtaking realizations, we assume that the overtaking probabilities are the same as for the highest measured WIP-level.

3 CASE DESCRIPTION

The re-entrant flow line shown in Figure 1 is used in the simulation study. Lots arrive at the network according to a Poisson process with mean interarrival time t_{a}. The re-entrant flow line consists of l identical workstations; each workstation consists of an infinite First-Come-First-Served (FCFS) buffer, and m identical parallel machines. The process time of the machines is gamma-distributed, with mean t_0 and coefficient of variation c_0. The number of times each lot is processed by the flow line is denoted r. The simulation case is modeled using the simulation language χ (Hofkamp and Rooda 2007). We refer to this simulation model as the 'detailed simulation model', to distinguish this model from the aggregate model.

To estimate $F_{\mathrm{E}}^{\mathrm{Q}}(t;w)$ and $F_{\mathrm{K}}^{\mathrm{Q}}(k;w)$, arrivals at and departures from the network are obtained from the detailed simulation model at a throughput ratio $\delta/\delta_{\max} = 0.8$, with $\delta = 1/t_{\mathrm{a}}$ being the actual throughput of the network and $\delta_{\max}$ the maximum obtainable throughput of the network. The algorithm in Appendix A is used to calculate EPT realizations and overtaking realizations, which are assigned to WIP-levels as explained in Section 2. We discard the first $3 \cdot 10^4$ EPT and overtaking realizations to account for the warm-up period. For each WIP-level w, $F_{\mathrm{E}}^{\mathrm{Q}}(t;w)$ and $F_{\mathrm{K}}^{\mathrm{Q}}(k;w)$ are estimated as explained in Section 2.

We test the aggregate model on two scenarios of the flow line. In Scenario I, the effect of the number of workstations l (the length of the line) on the prediction accuracy is investigated, which is studied for low, medium, and high variability of the process time distribution at the machines. In Scenario II, the effect of re-entrance is investigated.

To assess the accuracy of the cycle time predictions, we compare the cycle time distributions predicted by the aggregate model to the cycle time distributions obtained by the detailed simulation model of the network. The simulation model of the network is used to obtain the real cycle time distributions. For each scenario, we perform 10 simulation replications of 10^5 processed lots. For each replication run, the first $3 \cdot 10^4$ lots are discarded to account for the warm-up period. The same number of simulation replications, number of processed lots, and warm-up period are used in the simulation runs of the two aggregate models, which are also implemented in the language χ.

4 SCENARIO I: LENGTH OF THE FLOW LINE

Scenario I aims to investigate the effect of the number of workstations on the accuracy of the cycle time predictions by the aggregate models. The number of workstations $l = \{5, 10, 20\}$, and the coefficient of variability $c_0 = \{0.5, 1.0, 1.5\}$ are varied. The number of machines per workstations is fixed ($m = 10$) and there is no re-entrance ($r = 1$). The number of measured EPT and overtaking realizations after the warm-up period equals 10^6. We first present a selection of estimated aggregate model parameters. Next, the mean cycle times and cycle time distributions predicted by the aggregate model are presented and compared with the measured cycle times of the flow line being modeled.

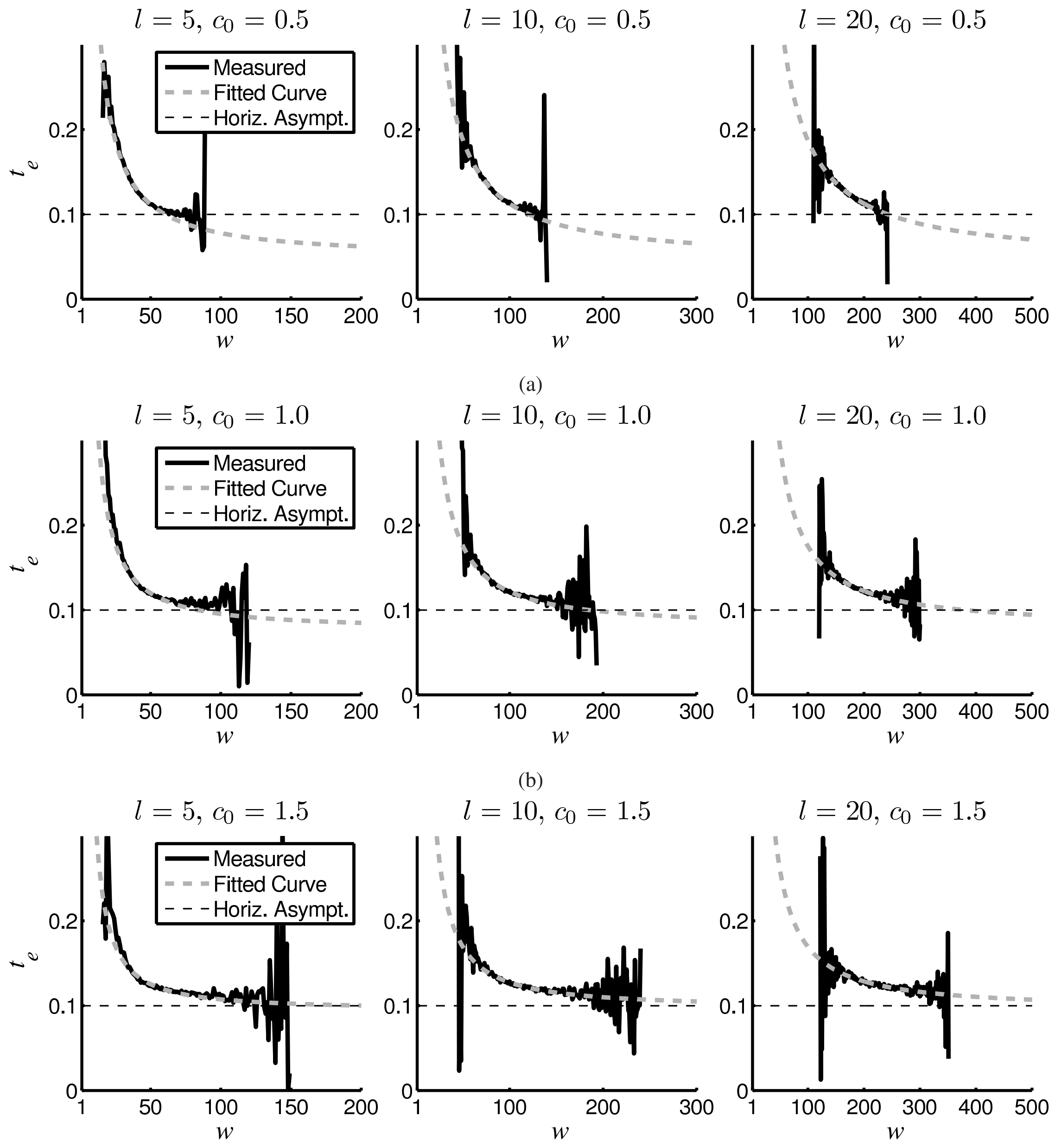
Estimated Aggregate Model Parameters

Figure 2 shows the measured mean EPT $t_{\mathrm{e},w}^{\mathrm{Q}}$ as a function of WIP level w (the black solid curves) and the corresponding fitted curves $\hat{t}_{\mathrm{e}}^{\mathrm{Q}}(w)$ (the grey dashed curves). The dashed black lines represent the expected horizontal asymptote, as determined by the maximum throughput. From left to right, the plots consider numbers of workstations $l = 5$, 10, and 20, respectively, with constant $m = 10$ and $r = 1$. Figure 2a considers $c_0 = 0.5$, Figure 2b $c_0 = 1.0$, and Figure 2c $c_0 = 1.5$.

We first discuss the results for the cases with $c_0 = 1.0$, shown in Figure 2b. The figure shows that for increasing l, the range of WIP levels for which EPT realizations are obtained becomes relatively smaller, and shifts to higher WIP levels. Noise in the $t_{\mathrm{e},w}^{\mathrm{Q}}$ estimates is clearly present, in particular at the edges of the range. In particular for $l = 10$ and $l = 20$, the reciprocal function (1) seems to be an appropriate function to fit the measured $t_{\mathrm{e},w}^{\mathrm{Q}}$ estimates in the range for which EPT realizations were obtained. In the aggregate model, the fitted curve is used for all WIP levels, including for WIP levels below and above the range for which EPT realizations were obtained. For increasing l, $t_{\mathrm{e},w}^{\mathrm{Q}}$ estimates have to be estimated by extrapolation of the fitted curve for a larger range of low and high WIP levels compared to the WIP range for which EPT realizations were obtained. As a consequence, for the high WIP levels, the horizontal asymptote predicted by the fitted curve is lower than the expected horizontal asymptote.

For $c_0 = 0.5$, Figure 2a shows that the horizontal asymptote predicted by the fitted curve is much lower than the asymptote that one would expect considering the maximum throughput of the system. For $c_0 = 1.5$, the predicted horizontal asymptote is very close to the expected one.

For the measured coefficient of variation of the EPT, $c_{\mathrm{e},w}^{\mathrm{Q}}$ (not shown here), we also observe that for increasing l, no estimates are measured for an increasing range of low and high WIP levels, similar to our observation for $t_{\mathrm{e},w}^{\mathrm{Q}}$. Consequently, these estimates also rely on the extrapolation of the fitted curve.



(a)

(b)

(c)

Figure 2: Measured mean EPT $t^{\mathrm{Q}}_{\mathrm{e},w}$ as a function of WIP level w, and the corresponding fitted curves for $l = \{5, 10, 20\}$ with $m = 10$, and $r = 1$; (a) considers $c_0 = 0.5$, (b) $c_0 = 1.0$, and (c) $c_0 = 1.5$.

Examination of the overtaking distribution (not shown here) shows that if l increases, the more lots may be overtaken, because there is on average more WIP in the system.

Cycle Time Predictions

Figure 3 shows the mean cycle time $\bar{\varphi}$ as a function of throughput ratio $\delta/\delta_{\max}$, measured at the network considered (the black solid curves), and predicted using the aggregate model (the grey dashed curves). From left to right, the plots show the CT-TH curves for $l = 5$, 10, and 20, with constant

$m = 10$, and $r = 1$. Figure 3a considers $c_0 = 0.5$, Figure 3b considers $c_0 = 1.0$, and Figure 3c considers $c_0 = 1.5$.

Figure 3 shows that for $c_0 = 1.0$, the throughput region for which accurate mean cycle time predictions are obtained becomes narrower for increasing l. This is because if the flow line becomes longer, the WIP range for which EPT realizations are obtained becomes relatively smaller (see Figure 2b). The estimates for the mean and coefficient of variation of the EPT at WIP levels higher or lower than the WIP range for which EPT realizations were obtained are predicted by the fitted curve. The fitted curve estimates become less accurate for WIP levels further away from the measured WIP range. We observe that the accuracy of the mean cycle time prediction particularly depends on the mean EPT estimates. We may therefore conclude that accurate extrapolation by the fitted curve for the mean EPT is crucial for the accuracy of the cycle time predictions, in particular for higher WIP levels than observed during the measurement period.

In the cases in which $c_0 = 0.5$, $\bar{\varphi}$ is underestimated for high throughput ratios. The reason is that the estimated horizontal asymptote for t_e is lower than the expected horizontal asymptote (see Figure 2a). For the case $c_0 = 1.5$ (Figure 3c), the mean cycle time is slightly underestimated for high throughput levels.

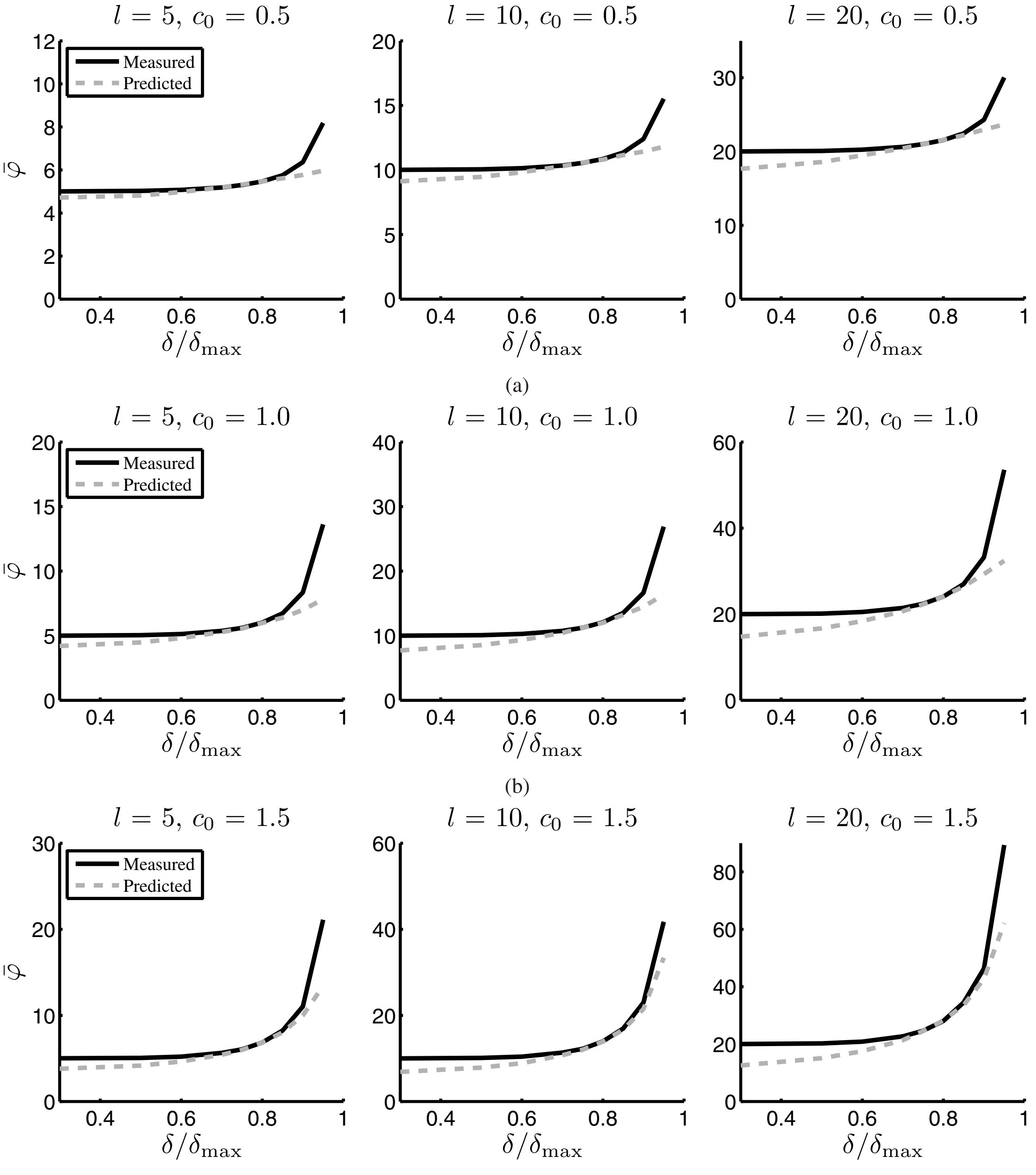
Figure 4 depicts the cycle time distributions measured at the network (the black curves), and the cycle time distributions predicted by the aggregate model (the dashed grey curves). The x-axes denote the cycle time φ, whereas the y-axes denote the probability $P(\varphi - \varepsilon < X < \varphi)$, where ε denotes the size of an interval, for which we choose 0.5. From left to right, the figure shows the cycle time distributions for throughput ratios of 0.6, 0.8, and 0.9, with constant $m = 10$, $c_0 = 1.0$, and $r = 1$. Recall that throughput ratio $\delta/\delta_{\max} = 0.80$ is the training level. The leftmost set of curves in each plot represents $l = 5$, the middle set $l = 10$, and the rightmost set $l = 20$.

Figure 4 shows that the cycle time distribution is predicted accurately at the training level ($\delta/\delta_{\max} = 0.8$). For other throughput ratios, the prediction accuracy deteriorates for increasing l. This was also observed for the mean cycle time (Figure 3b, $c_0 = 1.0$). Additionally, Figure 4 suggests that the aggregate model underestimates the variance of the cycle time distribution predicted for $\delta/\delta_{\max} = 0.60$, in particular for increasing l. This may be explained as follows: the WIP-dependent overtaking distribution is obtained at the training point of $\delta/\delta_{\max} = 0.80$. In the aggregate model this WIP-dependent overtaking distribution is used to predict the overtaking at higher or lower utilization levels. A higher utilization level means more WIP on average; a lower utilization level less WIP. If for a certain WIP-level no overtaking realizations were measured, it is assumed in the aggregate model that for this WIP level no overtaking occurs (as explained in Section 2). Increasing the length of the flow line (l) implies that the WIP range for which overtaking realizations are obtained becomes relatively smaller. As a result, for increasing l, it becomes more likely to encounter WIP levels for which no overtaking realizations were measured, when predicting the cycle time for throughput levels other than the training level. Consequently the amount of overtaking is underestimated, which causes the variability of the cycle time distribution to be underestimated as well.

For an accurate prediction of the cycle time distribution, a fitted curve for the overtaking distribution, which can also be used outside the interval of measured overtaking realizations, seems a necessary improvement. Similar to the fitted curves used in this paper for t_e and c_e, fitted curves might be used to estimate the mean and coefficient of variation of the number of overtaken lots as a function of w. The difference with the EPT is that the number of overtaking lots is a discrete variable that may realize values only between 0 and w (no more than w lots can be overtaken), whereas the EPT is a continuous variable that may exists between 0 and ∞. In Van Eenige (1996), families of fit functions are presented for discrete variables that realize values within a finite range. Those fit functions may be a starting point to fit the overtaking distribution.

5 SCENARIO II: RE-ENTRANCE

Scenario II aims to investigate the effect of re-entrance in the flow line on the accuracy of the cycle time predictions by the aggregate model. The number of times each lot is processed by the flow line $r = \{1, 2, 4\}$, and the process time variability $c_0 = \{0.5, 1.0, 1.5\}$ are varied. The length of the line, and the number of machines per workstations are fixed ($l = 5$, $m = 10$). The number of measured EPT and overtaking realizations after the warm-up period equals 10^6.



(a)

(b)

(c)

Figure 3: Mean cycle time φ as a function of throughput ratio $\delta/\delta_{\max}$ measured at the network, and predicted by the aggregate model, for $l = \{5, 10, 20\}$, with $m = 10$, and $r = 1$: (a) considers $c_0 = 0.5$, (b) $c_0 = 1.0$, and (c) $c_0 = 1.5$. Aggregate models are trained at $\delta/\delta_{\max} = 0.80$.

Estimated Aggregate Model Parameters

Figure 5 shows the measured $t^{\mathrm{Q}}_{\mathrm{e},w}$ values (the black solid curves) and the corresponding fitted curves $\hat{t}^{\mathrm{Q}}_{\mathrm{e}}(w)$ (the grey dashed curves) for $r = \{1, 2, 4\}$, with constant $l = 5$, $m = 10$, and $c_0 = 1.0$.

Figure 5 shows that the width of the WIP range, and the mean of the WIP range for which EPT realizations were measured are approximately the same for all values of r. This is because the

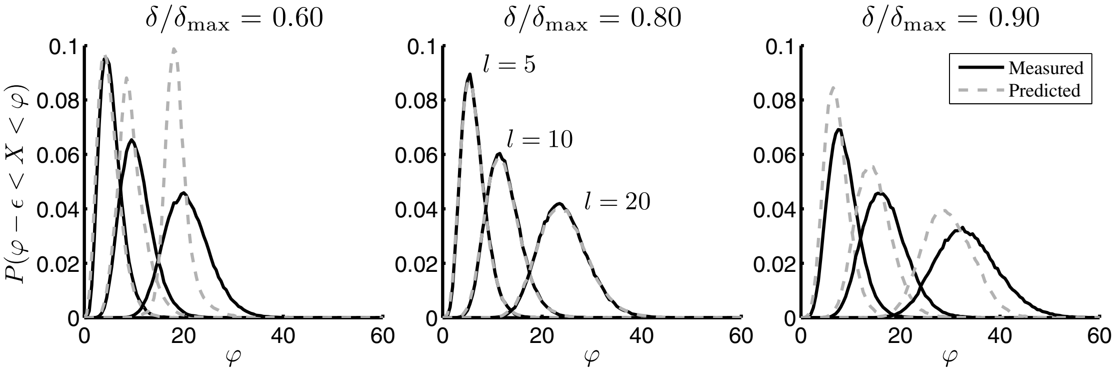


Figure 4: Cycle time distribution measured at the network, and predicted by the aggregate model for $l = \{5, 10, 20\}$, $m = 10$, $c_0 = 1.0$, and $r = 1$. Aggregate models are trained at $\delta/\delta_{\max} = 0.80$.

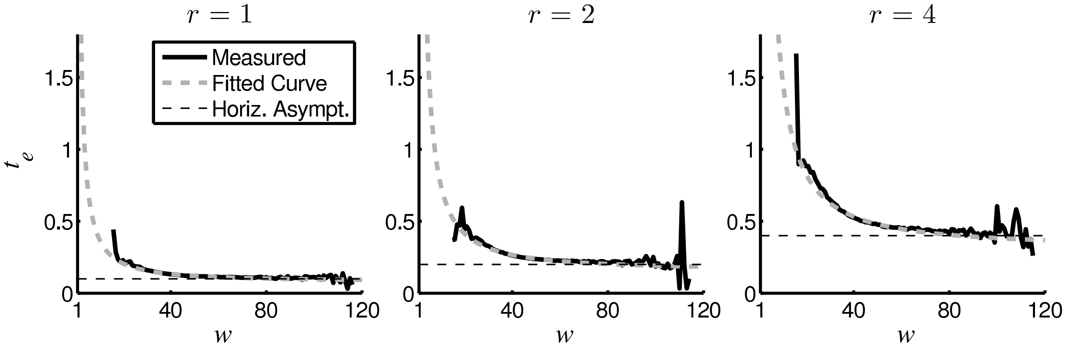


Figure 5: Measured mean EPT $t^{\mathrm{Q}}_{\mathrm{e},w}$ as a function of WIP level w, and the corresponding fitted curves for $r = \{1, 2, 4\}$ with $l = 5$, $m = 10$, and $c_0 = 1.0$.

throughput ratio $\delta/\delta_{\max}$ at which the arrival and departure events are measured is the same for all three values of r (which is achieved in the detailed simulation by multiplying mean inter-arrival time t_{a} by r). The mean and width of the observed WIP levels in the network depends on the number of machines and the variability in the network, respectively, which are the same for different values of r.

Examination of the measured overtaking probabilities shows that for increasing r, the maximum number of lots that may be overtaken decreases. The cause may be that for increasing r, the probability that an overtaking lot will be overtaken itself in a downstream workstation, or during a subsequent cycle, will increase.

Cycle Time Predictions

Figure 6 shows the CT-TH curves measured at the network considered (the black solid curves), and predicted by the aggregate model (the grey dashed curves). From left to right, the plots show the CT-TH curves for $r = 1$, 2, and 4, with constant $l = 5$, $m = 10$, and $c_0 = 1.0$.

Figure 6 shows that the accuracy of the mean cycle time predicted by the aggregate model is approximately the same for the different values of r. This is because the WIP range for which EPT realizations were obtained is also approximately the same for the different values of r (see Figure 5). Similar calculations were carried out to predict the cycle time for $c_0 = 0.5$ and $c_0 = 1.5$. These predictions also show that the accuracy of the cycle time distribution does not depend on r.

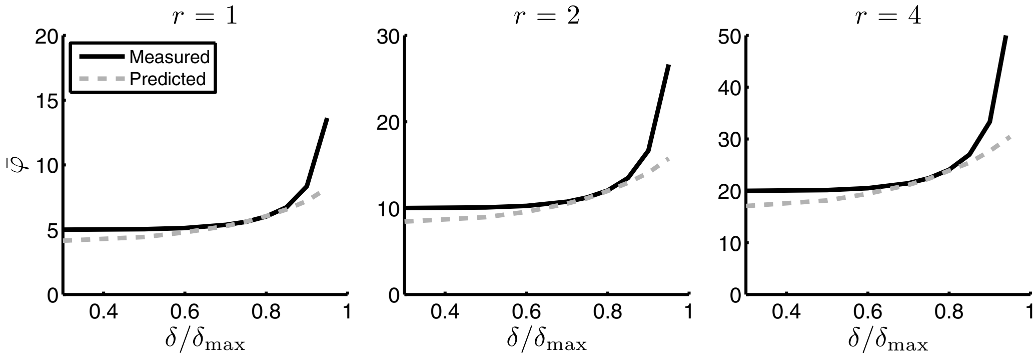


Figure 6: Mean cycle time φ as a function of throughput ratio $\delta/\delta_{\max}$ measured at the network, and predicted by the aggregate model, for $r = \{1,2,4\}$, with $l = 5$, $m = 10$, and $c_0 = 1.0$. Aggregate models are trained at $\delta/\delta_{\max} = 0.80$.

6 CONCLUSION

In this paper, we have investigated under which conditions the EPT-based aggregate modeling method presented in Veeger et al. (2009a) can accurately predict the cycle time distribution of a re-entrant flow line. The aggregate model was tested using a simulation case of a re-entrant flow line, motivated by semiconductor manufacturing, for which two different scenarios have been investigated.

In Scenario I, the length of the line (i.e., the number of workstations in the flow line) has been gradually increased. The accuracy of the prediction of the mean cycle time and variance of the cycle time deteriorates if the length of the flow line increases. The cause is that for increasing length of the flow line, EPT and overtaking realizations are measured in only a part of the WIP range for which the mean EPT varies. This measured part becomes smaller for increasing length of the flow line. As a consequence, the prediction of the fitted curve becomes less accurate, in particular for WIP levels further away from the WIP range for which EPT realizations were obtained. The accuracy of the predicted cycle times is particularly sensitive to the estimates of the mean EPT for high WIP levels, because they determine the maximum throughput of the system predicted by the aggregate model. This sensitivity increases if the WIP range for which EPT realizations are obtained becomes smaller. Regarding the overtaking probabilities, we assumed that for WIP levels lower than the WIP range in which we measured overtaking realizations, no overtaking occurs. This assumption reduces the prediction accuracy of the variance of the cycle time distribution at low throughput ratio compared to the training level, in particular for long flow lines.

In Scenario II, we have increased the amount of re-entrant cycles in the flow line. We have found that for both modeling approaches, the prediction accuracy is independent of the amount of re-entrant cycles.

The EPT-based aggregate model is considered particularly useful to make quick approximations of the cycle time for throughput levels relatively close to the working point, because it requires arrival and departure events at the network level only, and the aggregate model evaluations are computationally cheap.

There are a couple of questions that arise from the outcome of this work. The central question is whether it is possible to further improve the accuracy of the single-server aggregate model, in particular if the size of the aggregated system increases. Is there a theoretical limit to what can be reconstructed from information measured at a single operating point, to predict cycle times outside the operating point? If we need additional information from inside the network, what is the minimum additional information that is needed to arrive at a significantly more accurate single-server aggregation of a network? For instance, suppose that in a network the bottleneck station(s) are known. Can data (e.g., the maximum effective capacity) be used to correct the aggregate model regarding the maximum throughput prediction?

```
loop
    read $\tau, ev, i$
    if $ev = \mathbf{a}$ :
        if $\text{len}(xs) = 0$ :
            $(s, sw) := (\tau, 1)$
        end if
        $xs := xs \mathbin{+\!\!+} [(i, \text{len}(xs))]$
    elseif $ev = \mathbf{d}$ :
        write $\tau - s, sw$
        $(xs, k, aw) := \text{detOvert}(xs, i)$
        write $k, aw$
        if $\text{len}(xs) > 0$ :
            $(s, sw) := (\tau, \text{len}(xs))$
        end if
    end if
end loop
```

```
function $\text{detOvert}(xs, i)$
    $ys := []$
    while $\text{len}(xs) > 0$ :
        $(j, aw) := \text{head}(xs); xs := \text{tail}(xs)$
        if $j < i$ :
            $ys := ys \mathbin{+\!\!+} [(j, aw)]$
        elseif $j = i$ :
            return $(ys \mathbin{+\!\!+} xs, \text{len}(ys), aw)$
        end if
    end while
```

Figure 7: EPT Algorithm (left) and function detOvert (right)

A ALGORITHM

The algorithm used to calculate EPT-realizations and overtaking realizations (Veeger et al. 2009b) is depicted in Figure 7. The following variables are used: variable τ denotes the event time, variable ev the event type (arrival **a** or departure **d**), and i the lot arrival number (so lot i is the i^{th} arriving lot). Furthermore, variable xs is a list that contains for each lot in the system its arrival number, i, and the number of lots in the system upon its arrival, aw. Variable s is used to store the EPT start time. Variable sw denotes the number of lots in the system upon the EPT start. Variable k denotes the number of lots that a lot has overtaken. Function detOvert uses the following additional variables: ys is a list that stores part of list xs. Variable j stores a lot arrival number.

The EPT algorithm takes the aggregate model viewpoint. Upon an arrival event, a new EPT is started if the lot arrives in an empty system ($\text{len}(xs) = 0$). The start time s becomes τ and the corresponding wip-level is stored in variable sw. For every arriving lot, the lot arrival number i and the number of lots in the system upon arrival ($\text{len}(xs)$) are added to the end of list xs (indicated by $\mathbin{+\!\!+}$). When a departure event occurs, an EPT ends, the EPT being current time τ minus EPT start time s. The EPT is written to output along with number of lots in the system upon the EPT start sw. Next, the algorithm reconstructs how many lots k were overtaken by the departing lot using function detOvert, and furthermore returns number of lots aw in the system upon arrival of lot i and list xs with the information of lot i removed. The number of overtaken lots (k) and the number of lots in the system upon arrival of lot i (aw) are written. If there are still lots in the system after the departure ($\text{len}(xs) > 0$), a new EPT start time is stored in s, as well as the corresponding number of lots currently in the system ($\text{len}(xs)$).

The input of function detOvert consists of list xs and the arrival number i of the departing lot. The function iteratively removes each lot from xs and assigns its arrival number and the number of lots upon its arrival to variables j and aw respectively. If the arrival number of the observed lot is lower than the arrival number i of the departed lot, then (j, as) is concatenated to ys. If the arrival number j of the observed lot is equal to i, the function returns list $ys \mathbin{+\!\!+} xs$, which does not include lot i. Furthermore, the length of ys, and aw are returned. Note that the length of ys is equal to the number of lots that arrived earlier than lot i, but that are still in the system upon the departure of lot i. In other words, the length of ys is equal to the number of lots overtaken by lot i.

REFERENCES

Brooks, R. J., and A. M. Tobias. 2000. Simplification in the simulation of manufacturing systems. *International Journal of Production Research* 38 (5): 1009–1027.

Hofkamp, A., and J. Rooda. 2007. χ *1.0 reference manual*. Systems Engineering Group, Eindhoven University of Technology. <http://se.wtb.tue.nl/sewiki/chi/> [accessed March 31, 2010].

Johnson, R. T., J. W. Fowler, and G. T. Mackulak. 2005. A discrete event simulation model simplification technique. In *Proceedings 2005 Winter Simulation Conference*, ed. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 2172–2176. Orlando, Florida: Institute of Electrical and Electronics Engineers, Inc.

Kiba, J., S. Dauzère-Pérès, C. Yugma, and G. Lamiable. 2009. Simulation of a full 300mm semiconductor manufacturing plant with material handling constraints. In *Proceedings of the 2009 Winter Simulation Conference*, ed. M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and e. R. G. Ingalls, 1601–1609. Austin, Texas: Institute of Electrical and Electronics Engineers, Inc.

Miller, D. J. 1990. Simulation of a semiconductor manufacturing line. *Communications of the ACM* 33 (10): 98–108.

Rose, O. 2007. Improved simple simulation models for semiconductor wafer factories. In *Proceedings of the 2007 Winter Simulation Conference*, ed. S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 1708–1712. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Shanthikumar, J. G., S. Ding, and M. T. Zhang. 2007. Queueing theory for semiconductor manufacturing systems: a survey and open problems. *IEEE Transactions on Automation Science and Engineering* 4 (4): 513–522.

Van Eenige, M. J. A. 1996. *Queueing systems with periodic service*. Ph. D. thesis, Eindhoven University of Technology.

Veeger, C. P. L., L. F. P. Etman, J. van Herk, and J. E. Rooda. 2009a. Cycle time distributions of semiconductor workstations using aggregate modeling. In *Proceedings of the 2009 Winter Simulation Conference*, ed. M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and e. R. G. Ingalls, 1610–1621. Austin, Texas: Institute of Electrical and Electronics Engineers, Inc.

Veeger, C. P. L., L. F. P. Etman, J. van Herk, and J. E. Rooda. 2009b, May. Predicting the mean cycle time as a function of throughput and product mix for cluster tool workstations using ept-based aggregate modeling. In *Proceedings of the 2009 Advanced Semiconductor Manufacturing Conference (ASMC)*, 80–85. Berlin, Germany.

AUTHOR BIOGRAPHIES

CASPER VEEGER is a researcher in the Systems Engineering group of the department of Mechanical Engineering at the Eindhoven University of Technology. His research work is on the development of the effective process time method in semiconductor manufacturing. His email address is <c.p.l.veeger@tue.nl>.

PASCAL ETMAN is an associate professor in the Systems Engineering group of the department of Mechanical Engineering at the Eindhoven University of Technology. His research interests include simulation-based optimization, multidisciplinary design optimization, and the effective process time method for performance analysis of manufacturing systems. His email address is <l.f.p.etman@tue.nl>.

IVO ADAN is an associate professor in the department of Mathematics and Computer Science of the Eindhoven University of Technology. Since 2009, he also works as a part-time full professor at the Operations Research and Management group at the University of Amsterdam. His current research interests are in the analysis of multi-dimensional Markov processes and queueing models, and in the performance evaluation of communication, production and warehousing systems. His email address is <i.j.b.f.adan@tue.nl>.

JACOBUS ROODA is professor of the Systems Engineering group of the department of Mechanical Engineering at the Eindhoven University of Technology. His research interests include design and analysis of manufacturing systems, manufacturing control, and supervisory machine control. His email address is <j.e.rooda@tue.nl>.