Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

CYCLE TIME DISTRIBUTIONS OF SEMICONDUCTOR WORKSTATIONS USING AGGREGATE MODELING

Casper Veeger
Pascal Etman
Jacobus Rooda

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

Joost van Herk

NXP Semiconductors Nijmegen
Jonkerbosplein 52, FT3.109
6534AB Nijmegen
THE NETHERLANDS

ABSTRACT

Recently an aggregate modeling method has been developed to predict cycle time distributions as a function of throughput for manufacturing workstations with dispatching. The aggregate model is a single-server representation of the workstation with a workload-dependent process time distribution, and a workload-dependent overtaking distribution. The process time and overtaking distribution can be determined from arrival and departure events measured from the workstation at the factory floor. In this paper, we validate the proposed method in the context of semiconductor manufacturing. In particular we consider a lithography workstation. First, we present a simulation case that demonstrates the accuracy of the aggregate model to predict cycle time distributions. Second, we apply the aggregate modeling method to a case from semiconductor industry, and illustrate how the method performs using arrival and departure data obtained from the manufacturing execution system.

1 INTRODUCTION

Predicting cycle time distributions as a function of throughput is helpful in production planning for semiconductor workstations, such as the lithography workstation. The throughput is the number of lots processed per time unit. With cycle time, we mean the sum of queue time and process time of a lot at a specific workstation. From a cycle time distribution, quantiles can be derived. For instance, the 95% quantile defines the cycle time in which 95% of the lots are completed.

Only few analytical models to predict cycle time distributions of workstations exist. Queueing systems for which analytical models are available are e.g. the $M/M/1$ queue and the $M/M/m$ queue. Simulation is often the only option to calculate cycle time distributions. For example, Sivakumar and Chong (2001) used simulation to analyze cycle time distributions in semiconductor back-end manufacturing. Simulation-based analysis is computationally expensive. Yang, Ankenman, and Nelson (2008) therefore proposed to derive a metamodel from a detailed simulation model, which they use to derive cycle time quantiles as a function of the throughput.

The development of a detailed simulation model is usually time-consuming effort and it may be difficult to obtain all model parameters. Rose (2000) investigated the use of a simplified simulation model, in which the bottleneck workstation is modeled in detail and the remaining workstations in the network are lumped in a delay distribution. He concluded that the proposed model inaccurately estimates cycle time distributions for certain scenarios. Following up on this work, Rose (2007) introduced a utilization-dependent delay distribution determined by running a full detail simulation model at various utilization levels.

To avoid modeling the factory in full detail, in (Veeger, Etman, Lefeber, Adan, and Rooda 2009) an alternative simplified simulation model has been proposed. The proposed model is a single-server aggregate queueing model. The lumped parameters of the aggregate model can be determined without using a full-detail model but directly from measured arrival and departure events at the workstation at a single utilization level. The objective of the present paper is to demonstrate that the aggregate model can be effectively used to predict cycle time distributions of workstations in semiconductor manufacturing.

The Effective Process Time (EPT) concept is the basis of the aggregate model. It represents the aggregate process time distribution in the aggregate model. The EPT was originally defined by Hopp and Spearman (2008) as 'the time seen by a lot at a workstation from a logistical point of view'. The mean and variance of the EPT may be calculated from the raw process time and the various outage delay distributions in the process, and used in analytical equations representing the

$G/G/m$ system (Hopp and Spearman 2008). Data of the various distributions may not always be available. (Jacobs, Etman, van Campen, and Rooda 2003) therefore derived the EPT distribution parameters directly from arrivals and departures of lots at the workstation. In Kock, Etman, and Rooda (2008) this is cast into a EPT-based modeling framework. (Kock, Etman, Rooda, Adan, v. Vuuren, and Wierman 2008) proposed a $G/G/m$ alike aggregate model with workload-dependent EPT distributions, motivated by integrated processing types of machines which may have multiple lots in process at the same time. Both (Jacobs, Etman, van Campen, and Rooda 2003) and (Kock, Etman, Rooda, Adan, v. Vuuren, and Wierman 2008) developed EPT-based aggregate models that predict the mean cycle time as a function of the throughput. Due to the First-Come-First-Serve (FCFS) assumption in their aggregate model cycle time distributions are not accurately predicted.

The aggregate modeling method presented in (Veeger, Etman, Lefeber, Adan, and Rooda 2009) aims at predicting cycle time distributions. Similar to (Kock, Etman, Rooda, Adan, v. Vuuren, and Wierman 2008), the new aggregate model contains a workload-dependent EPT distribution, but additionally includes a probability distribution for lot overtaking. Like the EPT distribution, the overtaking distribution also depends on the number of lots in the system. The workload-dependent EPT distribution and overtaking distribution are determined from measured arrival and departure events. In (Veeger, Etman, Lefeber, Adan, and Rooda 2009), two simulation test examples are presented to demonstrate the potential of the new method.

In this paper, we validate the aggregate modeling method proposed in (Veeger, Etman, Lefeber, Adan, and Rooda 2009) on test cases motivated by semiconductor wafer fabrication. In particular we consider a workstation with lithography equipment, because this workstation is often the largest contributor to the cycle time. Lithography equipment is characterized by processing wafers of multiple lots at the same time (typically up to three lots). We present a simulation test case and a test case based on data obtained from the Crolles2 wafer factory. The simulation test case validates the proposed method and provides insight in the accuracy of the predicted cycle time distributions. The Crolles2 case demonstrates the applicability of the method in semiconductor practice.

The outline of the paper is as follows: the aggregate modeling method of (Veeger, Etman, Lefeber, Adan, and Rooda 2009) is explained in Section 2. The proposed method is validated with a simulation test case in Section 3, and the Crolles2 case is discussed in Section 4. Finally, we present conclusions in Section 5.

2 AGGREGATE MODELING METHOD WITH LOT OVERTAKING

Next we present the aggregation method for cycle time distribution prediction as first investigated in (Veeger, Etman, Lefeber, Adan, and Rooda 2009). We model a workstation as an infinitely buffered single-server aggregate queueing model with a workload-dependent process time distribution and a workload-dependent overtaking distribution. Figure 1a shows an example of a workstation, which consists of four parallel machines that each have three process steps. Between the first and second process step, there is a one-place buffer.

2.1 The Aggregate Model

Figure 1b visualizes the proposed aggregate model. Lots arrive in the system according to a Poisson process; lot i is defined as the i^{th} arriving lot in the system. The queue contains all w lots that are currently in the system. During service, lots stay in the queue (unlike common queue-server models). If the service time has elapsed, the lot that is currently first in the queue leaves the system. Upon arrival of a new lot i, it is determined how many lots in the queue are overtaken by lot i. The number of lots to overtake $k_i \in \{0,1,...,w\}$ is sampled from a probability distribution that depends on the amount of lots already in the system when lot i arrives (not including lot i). The arriving lot i is placed on position $w-k_i$ in the queue. For example, if $w=1$ upon arrival of lot i, there is a probability that no lots are overtaken ($k_i=0$), and a probability that one lot is overtaken ($k_i=1$). If no lots are overtaken, lot i is placed at the end of the queue (position $1-0=1$). If one lot is overtaken, lot i is placed ahead of the queue (position $1-1=0$).

Note that in the aggregate model, the server is not a true physical server; a timer determines when the next lot leaves the queue. The timer starts when: i) a lot arrives while no lots are present in the buffer, or ii) a lot departs while leaving one or more lots behind. When the timer starts, a time period is sampled from a probability distribution that depends on number of lots w in the system upon the timer start. The sampled time period is referred to as an Effective Process Time (EPT). When the EPT is finished, the first lot in the queue (position 0) leaves the system. While the EPT timer is not yet finished, new arriving lots may still overtake *all* lots in the system, including the first lot in the queue.

The input to the proposed aggregate model consists of an EPT distribution per wip (work in progress)-level and an overtaking probability function. We assume that the EPT-distributions are gamma distributed, and that the distributions for the various wip-levels are independent. We denote the overtaking probability function by $P(w,k)$, which is defined as the probability that k lots are overtaken given w lots in the system upon arrival.

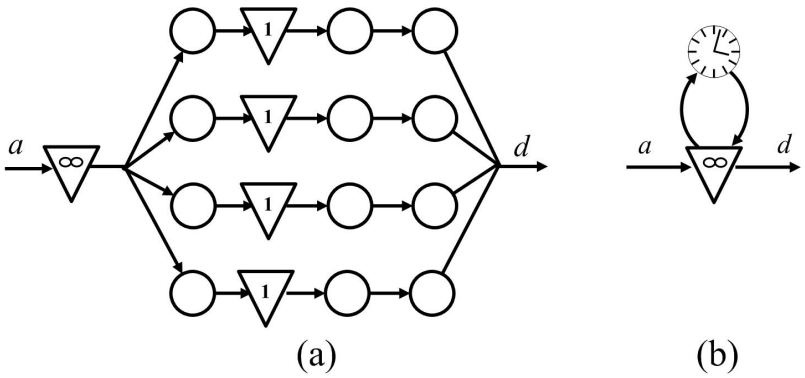

Figure 1: An example of a workstation (a), and the proposed aggregate model (b)

2.2 Calculating Model Parameters

To determine the EPT distributions and the overtaking probability function $P(w,k)$, arrival and departure data is measured from the workstation under consideration. For each lot i (which is the i^{th} *arriving* lot) departing from the workstation, departure time d_i is collected, as well as the corresponding arrival time a_i of the lot in the buffer of the workstation. From the arrival and departure data, the number of lots overtaken by each lot as well as the EPT realizations, are calculated using the algorithm shown Appendix A. The algorithm input consists of a lists of events; each event consists of time τ, event type *ev*, and lot arrival number i. The event type can be an arrival or a departure of a lot. The events are sorted in increasing time order.

The EPT algorithm takes the aggregate model viewpoint. The algorithm reconstructs the EPT realizations from the measured event list. A new EPT is started when i) an arrival event occurs while the system is empty, or ii) a departure event occurs while at least one lot remains in the system. An EPT ends when a departure event occurs. The algorithm then calculates the length of the EPT by subtracting the EPT start time from the departure event time τ. The EPT is written to output along with the number of lots w in the system upon the EPT start of lot i. Upon the departure of lot i, the algorithm also reconstructs how many lots (k) were overtaken by the departing lot i. This number is equal to the number of lots in the system upon departure of lot i that arrived earlier than lot i. The number of overtaken lots (k) and the number of lots w in the system upon arrival of lot i are written to output.

The EPT-realizations calculated by the algorithm are assigned to so-called buckets. Each bucket j corresponds to a number of lots w in the system *upon the EPT start*. A highest bucket N is defined, to which all EPT realizations are assigned that started when N or more lots were in the system. For each bucket j, the mean $t_{\text{e},j}$ and coefficient of variation $c_{\text{e},j}$ of the measured EPT distribution are determined, which are used in the process time gamma distributions of the aggregate model for the respective wip-levels. To obtain $P(w,k)$, overtaking realizations are also assigned to buckets, but the buckets now correspond to the number of lots in the system *upon arrival*.

2.3 Example

Suppose that we have two workstations that provide us with a list of events, visualized in Figure 2. Figure 2a shows four lots that do not overtake, and Figure 2b shows four lots with overtaking. We approximate both workstations with the proposed aggregate model.

The EPT realizations, and the number of lots overtaken by each arriving lot are depicted in Figure 2. For the EPT realizations, in between square brackets the number of lots in the system upon arrival is indicated. For the number of overtaken lots, in between the brackets the number of lots in the system upon arrival is indicated. In Figure 2a, Lot 1 arrives in an empty system and therefore an EPT is started. The number of lots in the system is 1. At time 5, the first departure occurs (that of Lot 1) and the EPT is ended. Since three lots remain in the system, a new EPT is started with $w = 3$. This EPT ends when Lot 2 departs at time 7, etc. Since the lots are processed in FCFS order, the number of overtaken lots is 0 for all four lots. Lot 1 arrives when no lots are in the system, Lot 2 arrives when there is one lot in the system, etc. In Figure 2b the EPTs are calculated in the same way. However, the number of overtaken lots is different. The number of lots

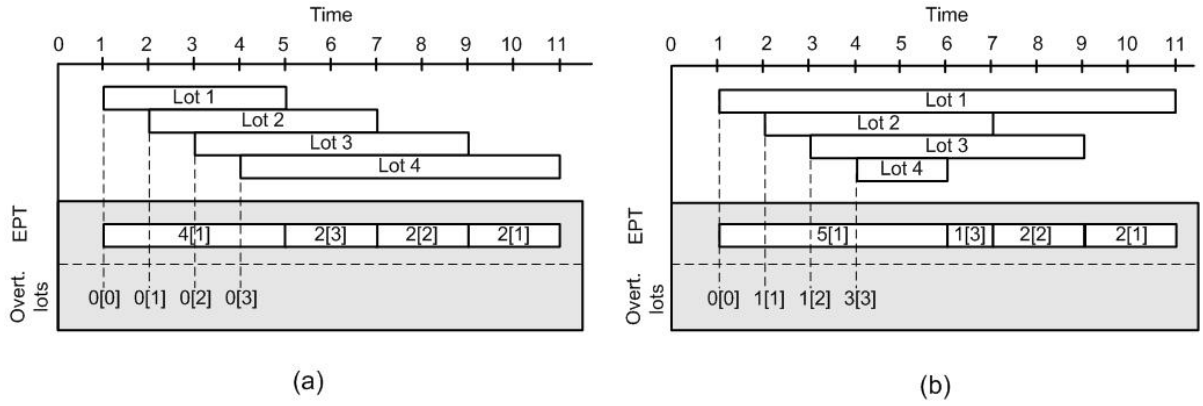

Figure 2: Lot-time diagrams of arrivals and departures including the calculated EPTs and number of overtaken lots; (a) without overtaking, (b) with overtaking

overtaken by Lot 1 is 0. Lots 2 and 3 overtake Lot 1 ($k = 1$), and Lot 4 overtakes Lot 1, 2 and 3 ($k = 3$). The number of lots upon arrival is the same as in Figure 2a.

3 Validation

A simulation test case is presented that represents a lithography workstation consisting of four track-scanner machines. The purpose of the simulation test case is to investigate the accuracy of the aggregate model representation in predicting cycle time distributions. Simulation results were obtained using the language χ (Hofkamp and Rooda 2007).

3.1 Description of the Case

The test case system is depicted in Figure 1a. Lots arrive at the infinite buffer according to a Poisson process: 50% of the arriving lots is of type A, whereas the other 50% is of type B. Lots are processed in First-Come-First-Serve order taking into account machine recipe qualification. If more than one qualified machine is available for processing, the lot is sent to the machine of which the first process has been idle longest (fairness). Each machine consists of three sequential process steps, with a one-place buffer between the first and second process. The first and third process step of each machine represent the track and have a constant process time of 1.0. The second process step represents the scanner.

Three scenarios are considered. In scenario 1, all machines are qualified for recipe A and B. The process time distribution of the second process step in each machine is exponential with a mean of 2.0. In scenario 2, the first machine is qualified only for recipe A, the second and third machine are qualified for recipe A and B, and the fourth machine is qualified only for recipe B. The second process step in each machine has a constant process time of 2.0. In scenario 3, the machine qualification is the same as in the second scenario; the process time distribution of the second process step is exponential with mean 2.0.

3.2 Calculating Model Parameters

For each test scenario, arrivals and departures of 10^6 lots were obtained at a throughput ratio of $\delta/\delta_{\max}$ of 0.8. Variable δ denotes the throughput, and $\delta_{\max}$ the maximum obtainable throughput of the system. We set maximum bucket number N to 20. The algorithm in Appendix A was used to calculate EPT realizations and overtaking realizations k, which were assigned to buckets as explained in the Section 2.2.

Figure 3 plots mean EPT t_e (left hand side) and coefficient of variability c_e of the EPT (right hand side) as a function of number of lots in the system w. The solid black line represents test scenario 1 (full qualification and exponential second process step). The dashed line represents the test scenario 2 (limited qualification and constant second process step). Finally, test scenario 3 (limited qualification and exponential second process step) is represented by the grey lines. For all test scenarios, t_e decreases for increasing w up to about $w = 16$, approaching 0.5. This is because for increasing w the system is

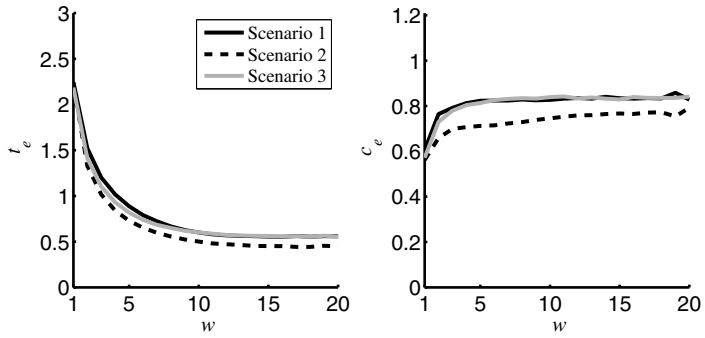

Figure 3: Mean EPT t_e and coefficient of variability c_e as a function of w

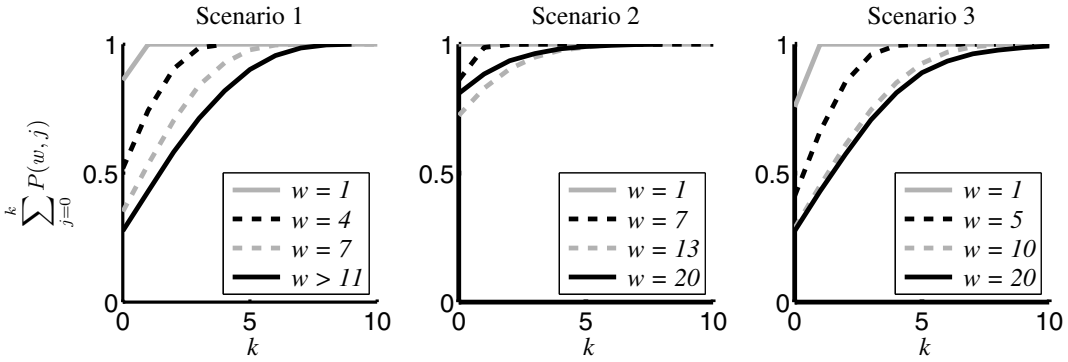

Figure 4: Cumulative overtaking probabilities as a function of w

more productive, since more process steps are processing lots. For test scenario 2, the system is slightly more productive (lower t_e) because less blocking occurs in the machines due to the constant process times of the servers. c_e approaches 0.8 for increasing w for scenario 1 and 3, and 0.7 for test scenario 2.

Figure 4 shows the cumulative overtaking probabilities $\sum_{j=0}^{k} P(w,j)$ as a function of k for several values of w. Recall that k is the number of overtaken lots, and w the number of lots in the system upon arrival. For test scenario 1, overtaking occurs due to parallel processing, because the process time in the second process step of each machine is exponential. Hence all w lots can be overtaken by an arriving lot with a maximum of 12 (when all process steps and the one-place buffer of three machines are occupied). In the test scenario 2, overtaking only occurs due to the dispatching rule (limited qualification of the servers) and not due to parallel processing because process times are constant. Hence, only queued lots of a different recipe than the arriving lot can be overtaken. In scenario 3, overtaking occurs due to parallel processing, and dispatching.

3.3 Cycle Time Predictions

Figure 5 depicts cumulative cycle time distributions for the three test scenarios for a throughput ratio $\delta/\delta_{\max}$ of 0.6, 0.8, and 0.9. Recall that the aggregate model parameters were obtained for $\delta/\delta_{\max} = 0.8$. The x-axis denotes the cycle time φ, the y-axis the cumulative probability $P(X \leq \varphi)$ that the cycle time is less than or equal to φ. Cycle times are obtained using simulation for 10^6 lots. The solid black lines represent the cumulative cycle time distributions of the test case system. The dashed grey lines give the cycle time distributions predicted by the aggregate model.

Figure 5 shows that the aggregate model generates remarkably accurate cycle distribution predictions, not only at the training level $\delta/\delta_{\max} = 0.8$, but also for $\delta/\delta_{\max} = 0.6$ and 0.9. Test scenario 2 is slightly less accurate than test scenario 1

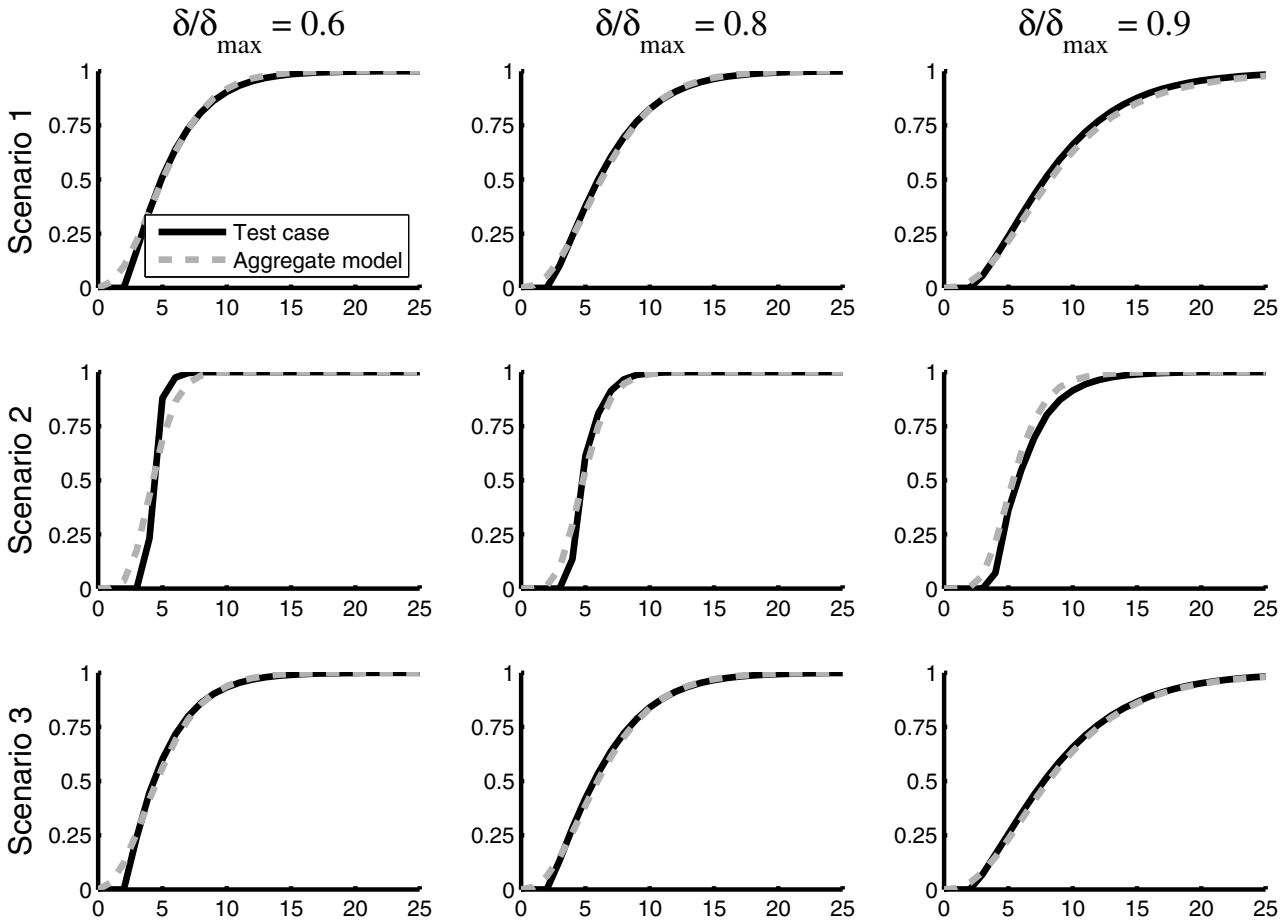

Figure 5: Cumulative cycle time distributions of the test case system, and the proposed aggregate modeling method. The x-axis denotes the cycle time, the y-axis denotes the cumulative probability.

and 3. The reason is the deterministic process time of all process steps. The only stochasticity is due to the arrival process. The accuracy of the aggregation seems to deteriorate the more deterministic behavior is observed.

3.4 Limited Amount of EPT-realizations

In a real manufacturing system, the number of arrival and departure events is usually much less than the 10^6 arrivals and departures we used in the simulation experiment. In this subsection, we limit the number of arrivals and departures to 20000, a number one may encounter in semiconductor manufacturing applications (see also Section 4). As a consequence, it is more difficult to accurately estimate model parameters t_e, c_e, and $P(w,k)$, and the cycle time predictions may deteriorate. In particular, we observe that an accurate estimation of t_e in bucket $j = N$ is crucial. This is because $1/t_{e,N}$ determines the predicted maximum throughput of the workstation. To maximize the accuracy of the $t_{e,N}$ estimation, we choose N as small as possible under the condition that $t_{e,j}$ is constant for $j \geq N$. Then, the number of EPT realizations in bucket N is the highest, while we do not discard the workload-dependency of t_e. Furthermore, for buckets $j < N$ we observe noise on $t_{e,j}$ and $c_{e,j}$ because little EPT realizations are available in each bucket. To overcome this problem of noise we introduce a curve fitting approach.

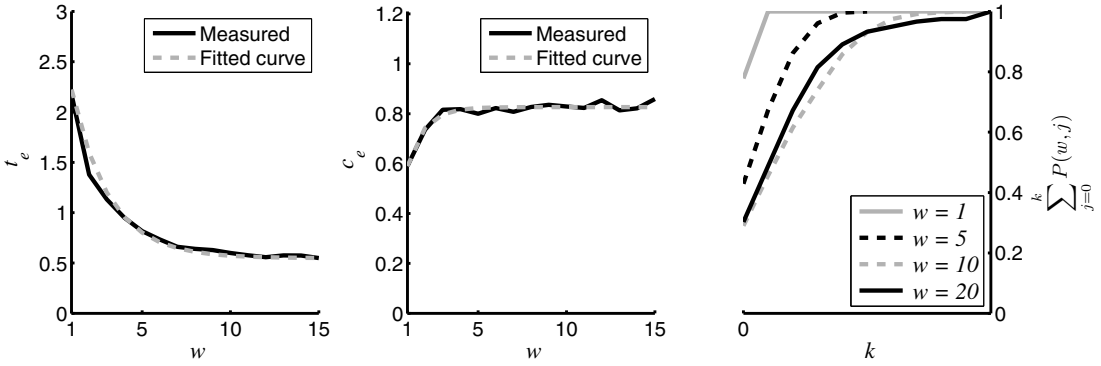

Figure 6: Measured mean EPT t_e (left) and coefficient of variability c_e (middle) with fitted curves, and measured cumulative overtaking probabilities (right) of test scenario 3 using 20000 arrivals and departures

The left plot of Figure 6 shows t_e as a function of w for test scenario 3 (the black line), but now obtained using only 20000 arrivals and departures. The middle plot of Figure 6 shows c_e as a function of w (the black line). We choose $N = 15$, because for buckets $j > 15$, t_e does not decrease further, as can be seen in the left plot of Figure 3. Note that Figure 6 now shows noise on the values of $t_{e,j}$ and $c_{e,j}$ for buckets $j < 15$, whereas this is not observed in Figure 3. To deal with the noise, we approximate $t_{e,j}$ by $\hat{t}_{e,j}$, for which we use the following exponential function (also used in (Veeger et al. 2008)):

$$\hat{t}_{e,j} = \theta + (\eta - \theta)e^{-\lambda(j-1)}. \tag{1}$$

Herein, θ represents the value of $\hat{t}_{e,j}$ at $j = \infty$. Variable η represents the value of $\hat{t}_{e,j}$ at $j = 1$. Variable λ represents the 'decay constant' of the exponential curve. We set η equal to the measured $t_{e,j}$ value in bucket $j = 1$. We set θ such that $\hat{t}_{e,N}$ is equal to the measured $t_{e,N}$. Variable λ is estimated using a non-linear least-squares fitting procedure. The values of θ, η, and λ we find are 2.224, 0.548, and 0.4716 respectively. We approximate $c_{e,j}$ by $\hat{c}_{e,j}$, and also use Equation (1). The values of θ, η, and λ obtained are 0.5899, 0.8265, and 1.0413 respectively.

The right plot of Figure 6 shows the cumulative overtaking probabilities $\sum_{j=0}^{k} P(w,j)$ as a function of k for several values of w. The overtaking probabilities are still sufficiently smooth to be used in the aggregate model. Hence, we do not introduce a curve fit here.

We use the aggregate model depicted in Figure 1b to estimate cycle time distributions of the test scenario 3, but we now use the fitted curves $\hat{t}_e$ and $\hat{c}_e$, and overtaking function $P(w,k)$ depicted in Figure 6 as model parameters (which were obtained using 20000 arrivals and departures). Cycle times predictions are obtained simulating the aggregate model for 10^6 lots.

Figure 7 depicts the cumulative cycle time distributions obtained for the considered workstation and the aggregate model for test scenario 3 at throughput ratios 0.6, 0.8, and 0.9. The figure shows that the accuracy is similar to the accuracy obtained using 10^6 arrivals and departures (as depicted in the bottom of Figure 5).

4 CROLLES2 CASE

We now apply the proposed method to a workstation in operation at the Crolles2 waferfab. Crolles2 is a multi-product 300mm fab in which both high volume products as well as small series and prototype products are produced. Standard production lots, so-called FOUPs (Front Opening Unified Pods), contain 25 wafers. In this section, we first describe the considered Crolles2 workstation, which is the lithography workstation. Subsequently, we explain how arrival and departure data was obtained and filtered. Next, we calculate from the arrival and departure data the EPT-distributions and overtaking probability function $P(w,k)$. Finally, cycle time distributions are predicted using the proposed aggregate model.

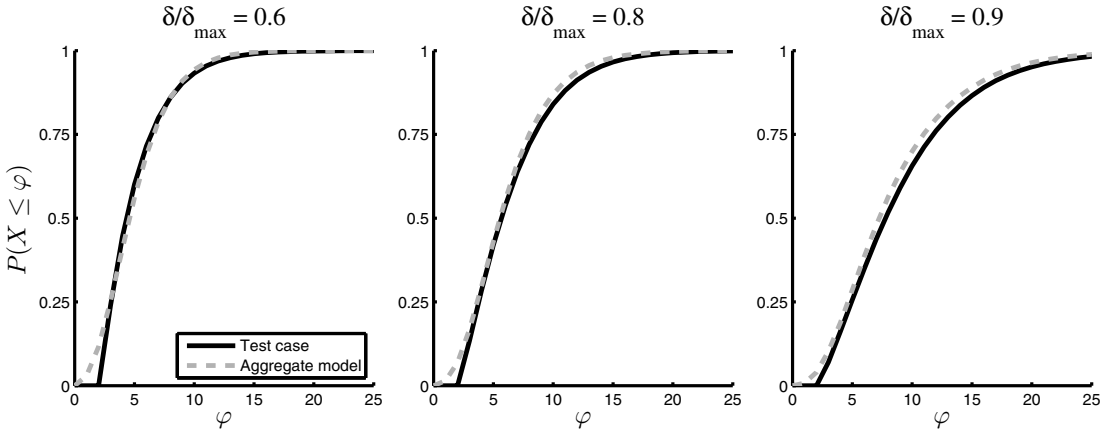

Figure 7: Cumulative cycle time distribution of test scenario 3, and predicted using 20000 arrivals and departures for throughput ratio 0.6, 0.8, and 0.9

4.1 Crolles2 Lithography Workstation

The lithography workstation consists of 14 track-scanner machines of different types, with different recipe qualifications. Lots are loaded on one of the load ports of a machine, after which wafers are sequentially loaded into the machine. First, wafers are cleaned, coated, and baked in the track. Then, the wafers are exposed in the scanner. Finally, the exposed wafers return to the track where they are developed and hard-baked. After all wafers of a lot have been loaded, the track starts loading the wafers of the next lot (if available on a load port). A track-scanner has four load ports; thus wafers of at most four lots can be in process at the same time, depending on the number of wafers per lot.

4.2 Estimating EPT-distribution Parameters

At the Crolles2 site, arrivals and departures of 42141 lots processed at the litho workstation were obtained from the Manufacturing Execution System (MES). To obtain arrivals and departures from MES data, the data processing algorithm described in (Veeger, Etman, van Herk, and Rooda 2008) is used. The EPT algorithm in Figure 10 is used to calculate EPT-realizations and lot overtaking realizations. We choose $N = 100$; for $j > 100$, t_e does not decrease further.

The left plot of Figure 8 shows the measured t_e values as a function of the number of lots w in the system upon the EPT start (the solid line). The middle plot depicts the measured c_e as a function of w. For reasons of confidentiality, no values on the y-axes are given. The dashed grey lines in the left and middle plot represents fitted curves, which we fit using the procedure described in Section 3 using exponential function (1). Note that we do not have EPT realizations for buckets $j < 18$. For t_e and c_e in bucket 1 we estimate values; t_e and c_e in buckets $1 < j < 18$ then follow from the curve fit.

The left plot of Figure 8 shows that the mean inter departure time decreases because the workstation becomes more productive for increasing w (more lots are in process), and approaches a minimum value for which the system works at its full throughput. The right plot of Figure 8 shows the measured cumulative overtaking probabilities $\sum_{j=0}^{k} P(w, j)$ as a function of k for several values of w. Note that for $w \geq 50$ considerable overtaking occurs.

4.3 Cycle Time Predictions

We use the aggregate model depicted in Figure 1b to estimate cycle time distributions of the lithography workstation, using the fitted curves $\hat{t}_e$ and $\hat{c}_e$, and overtaking function $P(w,k)$ depicted in Figure 8 as model parameters.

Figure 9 depicts cumulative cycle time distributions for the lithography workstation at relative throughput levels of 0.8, 0.9 and 1.0. The relative throughput is defined here as the throughput δ divided by the throughput at the working point δ^*. We use here the relative throughput instead of throughput ratio $\delta/\delta_{\max}$ because of confidentiality reasons. The rightmost plot represents the cumulative cycle time distribution at the working point of the workstation ($\delta/\delta^* = 1$). The x-axis denotes

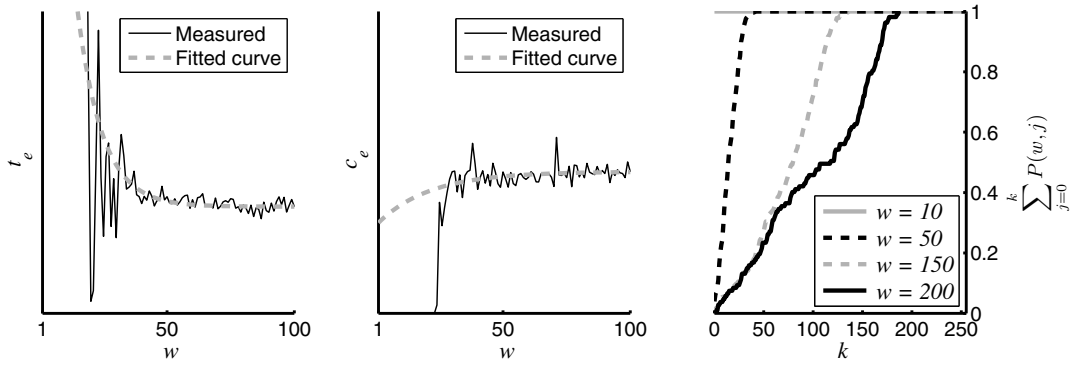

Figure 8: Measured mean EPT t_e (left) and coefficient of variability c_e (middle) with fitted curves, and measured cumulative overtaking probabilities (right) of the Crolles2 lithography workstation

cycle time φ, the y-axis cumulative probability $P(X \leq \varphi)$ that the cycle time is less than or equal to φ. The solid line in the rightmost plot represents the cycle time distribution of the workstation in the working point. The dashed lines represent the cycle time distributions estimated by the proposed method.

Figure 9 shows that the cycle time distribution is accurately estimated at the working point (the rightmost plot). For a decreasing relative throughput level, the predicted cycle times decrease. We can only verify the cycle time distribution at the working point. The simulation test case described in Section 3 indicates that accurate predictions can be made for throughput levels other than the working point, provided that the considered system is not predominantly deterministic. The Crolles2 lithography workstation is subject to stochastic behavior due to different recipes with different processing times, down behavior etc. Therefore, we expect that accurate cycle time distributions can be obtained at throughput levels other than the working point.

5 CONCLUSION

The investigated aggregate modeling method provides a simple and practical way to predict cycle time distributions for semiconductor workstations by means of simulation. The aggregate model is a single-server representation of the workstation that requires little development time. The process time in the model, which we refer to as the Effective Process Time (EPT), is sampled from a gamma distribution that depends on the momentary workload. Lots entering the buffer have a probability to overtake other lots according to a workload-dependent overtaking distribution. The EPT distribution and overtaking distribution are determined from arrival and departure events, measured at the workstation under consideration at a single utilization level.

The aggregate method has been validated in the context of semiconductor manufacturing. We have presented a simulation test case representing a lithography workstation, which shows that remarkably accurate cycle time distribution predictions can be obtained when we estimate the EPT distribution and overtaking distribution using 10^6 arrival and departure events. In a second experiment, we have used only 20000 arrival and departure events, which is an amount that is more reasonable in the context of semiconductor manufacturing. We have introduced a curve fitting approach to overcome the difficulties with noise that arise because of the limited amount of data. The accuracy of the prediction is in particular sensitive to the value of the parameter in the curve fit that represents the maximum throughput of the system.

We have demonstrated the applicability of the proposed method in semiconductor practice by applying the method on a the Crolles2 lithography workstation. We have obtained accurate cycle distribution predictions when comparing the simulated cycle time distributions with the measured cycle time distribution. The results of the simulation test case suggest that also accurate predictions can be made for throughput levels other than the operational throughput.

The results obtained in this paper imply that the proposed method can be used to make a trade-off between throughput and cycle time for the lithography workstation. For example, the maximum throughput can be estimated for which 95% of the lots are completed within a user-defined time span. Lithography is usually the main contributor to the cycle time of lots. However, we may expect that the proposed method can also be used for other semiconductor workstations, such as the metal

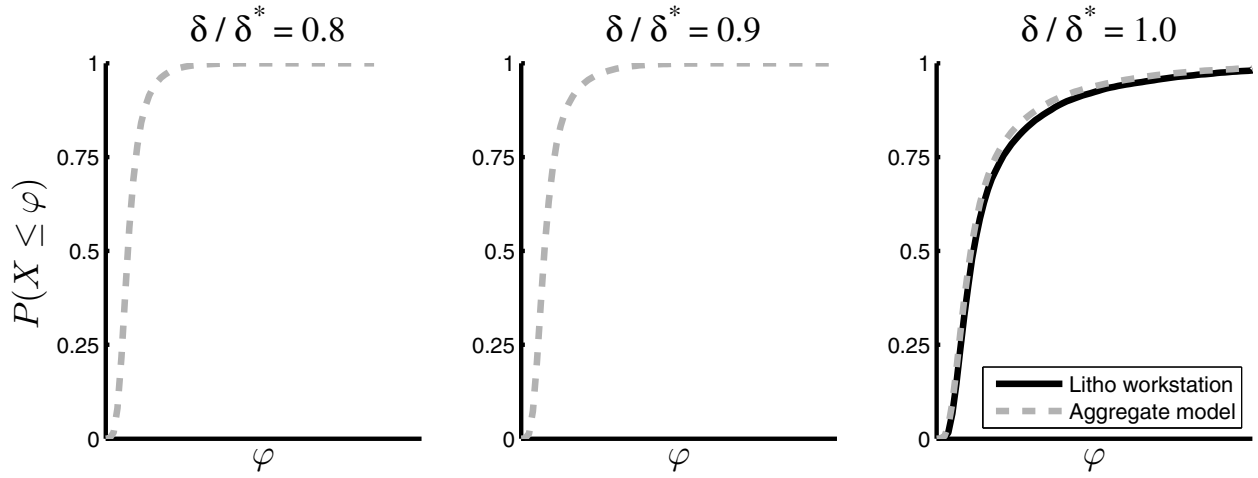

Figure 9: Cumulative cycle time distribution of the litho workstation, and predicted by the proposed method for relative throughput levels of 0.8, 0.9, and 1.0

or implant workstations. These workstations may also have wafers of multiple lots in process at the same time. In future research we want to investigate whether we can aggregate entire manufacturing networks into a single-server representation.

ACKNOWLEDGMENTS

We thank Bart Lemmen of Crolles2 for his support in obtaining the data.

A ALGORITHM

The algorithm used to calculate EPT-realizations and overtaking realizations (Veeger et al. 2009) is depicted in Figure 10. 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 $+\!\!+$). 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 $ys +\!\!+ xs$, 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.

```
loop
    read τ, ev, i
    if ev = a :
        if len(xs) = 0 :
            (s, sw) := (τ, 1)
        end if
        xs := xs ++ [(i, len(xs))]
    elseif ev = d :
        write τ − s, sw
        (xs, k, aw) := detOvert(xs, i)
        write k, aw
        if len(xs) > 0 :
            (s, sw) := (τ, len(xs))
        end if
    end if
end loop
```

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

Figure 10: EPT Algorithm (left) and function detOvert (right)

REFERENCES

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 August 20, 2009].

Hopp, W. J., and M. L. Spearman. 2008. *Factory physics: Foundations of manufacturing management*. third ed. New York: IRWIN/McGraw-Hill.

Jacobs, J. H., L. F. P. Etman, E. J. J. van Campen, and J. E. Rooda. 2003, Aug. Characterization of operational time variability using effective process times. *IEEE Transactions on semiconductor manufacturing* 16 (3): 511–520.

Kock, A. A. A., L. F. P. Etman, and J. E. Rooda. 2008. Effective process times for multi-server flowlines with finite buffers. *IIE Transactions* 40 (3): 177–186.

Kock, A. A. A., L. F. P. Etman, J. E. Rooda, I. J. B. F. Adan, M. v. Vuuren, and A. Wierman. 2008, August. Aggregate modeling of multi-processing workstations. Eurandom report, <http://www.eurandom.tue.nl/reports> [accessed August 20, 2009].

Rose, O. 2000. Why do simple wafer fab models fail in certain scenarios? In *Proceedings of the 2000 Winter Simulation Conference*, ed. J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 1481–1490. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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.

Sivakumar, A. I., and C. S. Chong. 2001. A simulation based analysis of cycle time distribution, and throughput in semiconductor backend manufacturing. *Computers in Industry* 45 (1): 59–78.

Veeger, C. P. L., L. F. P. Etman, A. A. J. Lefeber, I. J. B. F. Adan, and J. E. Rooda. 2009, June. Predicting flow time distributions in workstations with dispatching: an aggregate modeling approach. In *Proceedings Stochastic Models of Manufacturing and Service Operations (SMMSO)*, 38–45.

Veeger, C. P. L., L. F. P. Etman, J. van Herk, and J. E. Rooda. 2008, May. Generating cycle time-throughput curves using effective process time based aggregate modeling. In *Proceedings of the 2008 Advanced Semiconductor Manufacturing Conference (ASMC)*, 127–133.

Yang, F., B. E. Ankenman, and B. L. Nelson. 2008, Fall. Estimating cycle time percentile curves for manufacturing systems via simulation. *INFORMS Journal on Computing* 20 (4): 628–643.

AUTHOR BIOGRAPHIES

CASPER VEEGER is a Ph.D. student 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 for these proceedings is <c.p.l.veeger@tue.nl>.

PASCAL ETMAN is an assistant 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 for these proceedings is <l.f.p.etman@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 for these proceedings is <j.e.rooda@tue.nl>.

JOOST VAN HERK is working as quality assurance and safe launch engineer in the Business Line Automotive Safety and Comfort at NXP Semiconductors Nijmegen. His interests include the optimization of the quality of products, manufacturing effectiveness, and advanced equipment and process control. His email address for these proceedings is <joost.van.herk@nxp.com>.