Predicting Cycle Time Distributions for Integrated Processing Workstations: An Aggregate Modeling Approach

C. P. L. Veeger, L. F. P. Etman, E. Lefeber, I. J. B. F. Adan, J. van Herk, and J. E. Rooda

***Abstract*—To predict cycle time distributions of integrated processing workstations, detailed simulation models are almost exclusively used; these models require considerable development and maintenance effort. As an alternative, we propose an aggregate model that is a lumped-parameter representation of the workstation. The aggregate model is a single server with a work-in-process dependent aggregate process time distribution and overtaking distribution. The lumped parameters are determined directly from arrival and departure events measured at the workstation. An extensive simulation study and an industry case demonstrate that the aggregate model can accurately predict the cycle time distribution of integrated processing workstations in semiconductor manufacturing.**

***Index Terms*—Cycle time distribution, discrete-event simulation, factory dynamics, manufacturing systems, performance evaluation, queueing.**

I. Introduction

IN PRODUCTION planning for semiconductor workstations, there is a tradeoff between productivity and the cycle time. A workstation consists of a group of machines that perform similar operations, and that share the same input buffer. Workstation productivity is expressed as the number of lots processed per time unit, which is also referred to as throughput. High productivity is desirable due to the capital intensive equipment used. On the other hand, high productivity causes long "cycle times," defined as the sum of the process time and the waiting time at the workstation. High cycle times may negatively influence the on-time delivery performance of the semiconductor manufacturing system, or the time-to-market of new products.

Manuscript received November 27, 2009; revised April 9, 2010 and November 11, 2010; accepted November 13, 2010. Date of publication December 3, 2010; date of current version May 4, 2011.

C. P. L. Veeger, L. F. P. Etman, E. Lefeber, and J. E. Rooda are with the Systems Engineering Group, Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven 5600 MB, The Netherlands (e-mail: c.p.l.veeger@tue.nl; l.f.p.etman@tue.nl; a.a.j.lefeber@tue.nl; j.e.rooda@tue.nl).

I. J. B. F. Adan is with the Stochastic Operations Research Group, Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven 5600 MB, The Netherlands, and also with the Operations Research and Management Group, Department of Quantitative Economics, University of Amsterdam, Amsterdam 1012, The Netherlands (e-mail: i.j.b.f.adan@tue.nl).

J. van Herk is with NXP Semiconductors, Nijmegen 6534 AE, The Netherlands (e-mail: joost.van.herk@nxp.com).

Digital Object Identifier 10.1109/TSM.2010.2094630

To make a tradeoff between productivity and cycle times, an accurate prediction of the cycle time distribution as a function of the throughput is required. For this prediction, a model may be used that has to incorporate semiconductor workstation behavior such as integrated processing, outage delays, and dispatching rules. Integrated processing machines can process multiple lots at the same time in the various process chambers. Examples of integrated processing machines are lithography machines, and cluster tools. For planning purposes it is desirable that the model requires little development and maintenance effort, and that model evaluations are computationally cheap.

To predict cycle time distributions, simulation models are almost exclusively used. Application of classical queueing models, such as the $G/G/m$ queue [1], is mostly restricted to relatively simple systems, and implementation in the semiconductor industry has been unsatisfactory [2]. Alternatively, statistical analysis of historical data (e.g., data mining) may be used to predict future expected cycle times [3]–[6], but these approaches do not focus on cycle time distribution prediction.

Predictions of the cycle time distribution may be obtained using a detailed simulation model. For example, [7] and [8] estimated a set of quantiles from a detailed simulation model by processing simulation output using a Cornish-Fisher expansion. Sivakumar and Chong [9] used a detailed simulation model to analyze cycle time distributions in semiconductor back-end manufacturing. Detailed simulation models allow the inclusion of many details of the factory floor to arrive at accurate predictions, and can be easily updated if the factory conditions change (e.g., if an additional machine is installed). On the other hand, detailed models are computationally expensive. Dangelmayer *et al.* [10] pointed out that model abstraction is necessary to allow simulation experiments of efficient runtime.

One way to make an abstraction of a detailed simulation model is to carry out simulation runs according to a design of experiments, and use the responses to generate a metamodel. For example, Yang *et al.* [11] and Chen [12] built a metamodel from a detailed simulation model, which they used to derive cycle time quantiles as a function of the throughput.

Another approach to abstract a detailed simulation model is aggregation. Brooks and Tobias [13], and Johnson *et al.*

[14] used a simplification technique in which non-bottleneck workstations are replaced by a constant delay, but they do not use their simplified model for cycle time distribution prediction. Rose [15] used delay distributions to aggregate all workstations except the bottleneck station. He concluded that the proposed model inaccurately estimates cycle time distributions for certain scenarios. To improve the cycle time estimations, Rose [16] replaced the delay distributions by a first-come-first-served (FCFS) single-server system with inventory-dependent process times, which are determined by running a full-detail simulation model at various utilization levels.

Model abstraction techniques as described above require that a detailed simulation model is available beforehand. Development of such a detailed simulation model requires substantial resources to develop and maintain [2].

In this paper, we also propose an aggregate model, but we do not need to model the system in full detail first. Unlike [13]–[16], we consider single workstations in this paper, instead of flow lines of workstations. Our intention is that such aggregate workstation representations can also be used as a building block in a model of the entire factory. This paper proposes a new single-server aggregate queueing model for integrated processing workstations in semiconductor manufacturing. The lumped parameters of the model are determined from lot arrival times and lot departure times, measured at the workstation in operation. We refer to the average throughput level of the workstation during the measurement period as the "training level." We demonstrate that the aggregate model can accurately predict cycle time distributions of workstations in semiconductor manufacturing, also for throughput levels *other* than the training level.

The process time distributions and outage delays in the workstation are aggregated by means of a work in process (WIP)-dependent aggregate process time distribution. By WIP we mean the total number of lots in the workstation including the input buffer. We refer to the aggregate process time as the effective process time (EPT). The EPT was introduced by Hopp and Spearman [17], who defined the EPT as "the process time seen by a lot at a workstation." They calculated the mean and the variance of the EPT from the raw process time, and the preemptive and non-preemptive outages. They used the mean and variance of the EPT in closed-form $G/G/m$ equations to predict the mean cycle time. Because data of the various distributions may not always be available, Jacobs *et al.* [18] developed an algorithm to determine the EPT distribution parameters directly from arrivals and departures measured at the workstation.

For semiconductor workstations, the EPT-distribution parameters are typically WIP dependent, because wafers of multiple lots may be in process at the same time. In this paper, we consider workstations with cascading machines, in which the process times of multiple lots partially overlap (e.g., a lithography workstation and workstations with cluster tools). We do not consider workstations with batching machines. WIP-dependency of the EPT distribution parameters can also be caused by outage delays that may occur when the machine is idle [19], such as preventive maintenance. The attribution of such delays to the EPT may be utilization-dependent [19]. Therefore, Kock *et al.* [20] proposed a $G/G/m$-like aggregate simulation model with a WIP-dependent EPT-distribution to predict the *mean cycle time*. Veeger *et al.* [21] demonstrated that the method of [20] is able to predict the mean cycle time as a function of the throughput for workstations in an operating semiconductor environment. However, the aggregate model of [20] does not necessarily yield accurate *cycle time distribution* predictions, due to the FCFS rule in the aggregate model.

In this paper, we use a WIP-dependent EPT distribution similar to [20], but additionally take into account *the order* in which lots are processed. Each lot that arrives in the aggregate model has a probability to overtake a number of other lots already in the system. The number of lots to overtake is determined by a WIP-dependent overtaking distribution. Like the EPT distribution, the lot overtaking distribution is determined from measured arrival and departure events.

We demonstrate that the proposed method can quite accurately predict cycle time distributions for semiconductor workstations. We first validate the method using a simulation test case of a workstation where we vary the number of parallel machines, the number of integrated processes, the dispatching rule, and the variability of the process time and the interarrival time. In this simulation case, sufficient arrival and departure events are available to accurately estimate the EPT and overtaking distribution. However, in semiconductor practice, typically a limited number of measured events is available. In a second simulation case, motivated by a lithography workstation, we show how accurate predictions can still be made when a limited amount of data is available. We also use the second case to investigate the prediction accuracy when two different product types are produced. Finally, a test case based on data from the Crolles2 wafer factory in Crolles, France, demonstrates the applicability of the method in semiconductor manufacturing practice.

The outline of this paper is as follows. The proposed aggregate modeling method is explained in Section II. The validation experiments are presented in Section III, and the Crolles2 case is discussed in Section IV. Finally, we present our conclusions in Section V.

II. Model Concept

We model a workstation as an infinitely buffered single-server aggregate queueing model with a WIP-dependent process time distribution and a WIP-dependent overtaking distribution. Fig. 1(a) illustrates a workstation consisting of m parallel flow lines in which l lots can be processed simultaneously. Each flow line may represent an integrated processing machine, such as a lithography track-scanner cell, which may process wafers of up to four lots at the same time. Another example is a cluster tool, which may typically process wafers of up to two lots at the same time. Fig. 1(b) visualizes the proposed aggregate model. In this section, we introduce the aggregate model concept and explain how we determine model parameters.

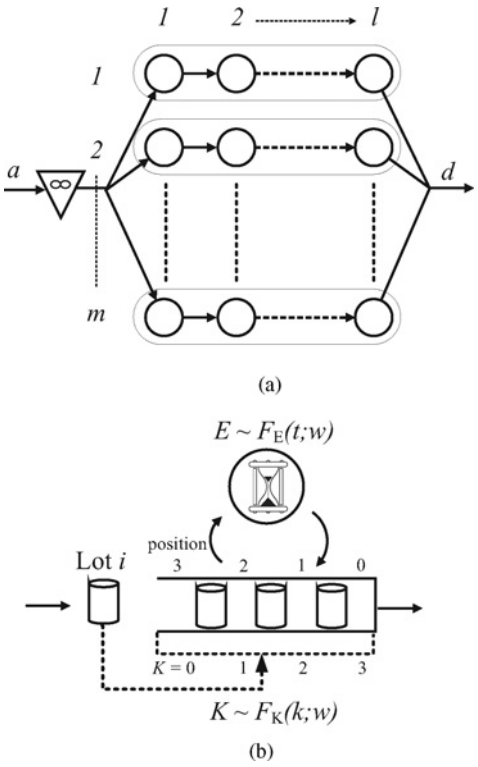

Fig. 1. (a) Example of a workstation. (b) Proposed aggregate model.

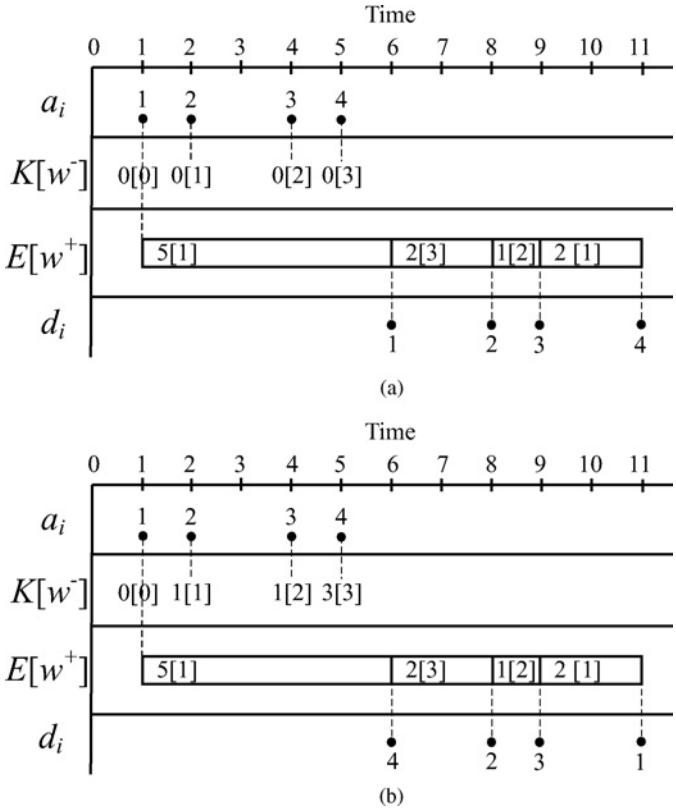

Fig. 2. Lot-time diagrams of four lots processed by the aggregate model including the EPTs sampled by the timer and the sampled number of overtaken lots. (a) Without overtaking. (b) With overtaking.

A. Aggregate Model

We propose the following aggregate model [Fig. 1(b)]. Note that the structure of the aggregate model differs significantly from the real workstation. Lots arrive in the queue of the aggregate model according to some arrival process. Lot i is defined as the ith *arriving* lot in the queue. The queue is *not* a queue as in common queue-server models (such as the $G/G/1$ model), but contains *all* lots that are currently in the system including the lots that are supposed to be in process. So during the process, lots stay in this queue. If the process 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 already present in the queue w will be overtaken by Lot i. The number of lots to overtake $K \in \{0, 1, ..., w\}$ is sampled from probability distribution $F_K(k; w)$, which defines the probability $P(K \leq k; w)$ that at most k lots are overtaken. Probability distribution $F_K(k; w)$ depends on the number of lots w in the queue just before Lot i arrives (so not including Lot i itself). The arriving Lot i is placed on position $w - K$ in the queue, where position 0 is the head of the queue. For example, in Fig. 1(b), $w = 3$ upon arrival of Lot i. In this case, there is a probability that 0, 1, 2, or 3 lots will be overtaken ($K = 0$, 1, 2, or 3). If no lots are overtaken, Lot i is placed at the end of the queue (position $3 - 0 = 3$). If one lot is overtaken, Lot i is placed after the first two lots in the queue, and before the last lot in the queue (position $3 - 1 = 2$), and so on.

We emphasize that in the aggregate model, the server is not a true physical server, but a timer that determines when the next lot leaves the queue. We model the server as a timer to allow newly arriving lots to overtake *all* lots in the system while the timer is running. The timer starts when: 1) a lot arrives while no lots are present in the queue or 2) a lot departs while leaving one or more lots behind. When the timer starts, a time period E is sampled from probability distribution $F_E(t; w)$, which defines the probability $P(E \leq t; w)$ that E is less than or equal to t. The probability distribution $F_E(t; w)$ depends on number of lots w in the system just after the timer start. So in case of a lot arrival (case 1), w includes the lot that just arrived. In case of a lot departure (case 2), w does not include the departed lot. Time period E is referred to as an EPT. When the EPT is finished, the lot that is presently first in the queue (position 0) leaves the system.

The input of the aggregate model consists of an EPT distribution $F_E(t; w)$ per WIP-level w and an overtaking distribution $F_K(k; w)$ per WIP-level w. We assume that all sampled EPT realizations, and overtaking realizations are independent and identically distributed within the same WIP-level.

B. Example

Fig. 2(a) shows four lots processed by the aggregate model in FCFS order. The first row of Fig. 2(a) shows the arrivals a_i of each Lot i (i indicates the arrival number). The second row depicts the numbers of overtaken lots K, which are sampled—upon each lot arrival—from the overtaking probability distribution corresponding to number of lots in the queue w^- (depicted in between square brackets); we use w^- in Fig. 2(a) instead of w to point out that we mean here the WIP just *before* the arrival of Lot i, not including Lot i. The third row

in Fig. 2(a) depicts the EPT realizations E, which are sampled upon each EPT start by the timer from the EPT distribution corresponding to number of lots in the queue w^+ (depicted in between square brackets); w^+ indicates the WIP just *after* the event (an arrival or a departure) that triggered the EPT start. The fourth row depicts the resulting departures d_i. Fig. 2(a) shows that for each arrival the sampled number of overtaken lots equals zero, which implies that no overtaking occurs, so the order of arrival is equal to the order of departure.

Fig. 2(b) shows four lots with overtaking. The lot arrival times, and the sampled EPTs are the same as in Fig. 2(a), but the sampled values of K are different. Upon arrival of Lot 2, K becomes 1 so Lot 2 overtakes one lot (Lot 1). Lot 3 also overtakes one lot (Lot 1 again), and Lot 4 overtakes three lots (Lots 1, 2, and 3). So when the timer first ends, Lot 4 is ahead of the queue and departs. Next Lot 2 departs, then Lot 3, and then Lot 1.

C. Calculating Model Parameters

To determine EPT distribution $F_{\mathrm{E}}(t; w)$ and overtaking distribution $F_{\mathrm{K}}(k; w)$, the aggregate model is trained using arrival and departure data measured at the workstation under consideration. For each Lot i (which is the ith *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, we determine the EPT realizations, the number of lots overtaken by each lot, as well as the corresponding WIP-levels using the algorithm given in 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 keeps track of the momentary WIP-level and reconstructs the EPT realizations from the measured event list. A new EPT is started when: 1) an arrival event occurs while the system is empty or 2) 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 duration of the EPT by subtracting the EPT start time from the departure time (event time τ). The EPT is written to an output file 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. A lot has been overtaken by Lot i when it arrived earlier than Lot i (so has a lower arrival number i), but departs later than Lot i. Hence, the value of k is calculated by counting the number of lots still in the system upon departure of Lot i that have a lower arrival number lower than i. The number of overtaken lots k and the number of lots w in the system upon arrival of Lot i are written to an output file.

The EPT-realizations calculated by the algorithm are grouped according to the number of lots w in the system upon the EPT start. For implementation reasons, we define a maximum WIP-level $w_{\max}$, in which all EPT realizations are grouped that started with $w \geq w_{\max}$ lots in the system. For each WIP-level w, we obtain a distribution, which is used in the aggregate model for the EPT distribution $F_{\mathrm{E}}(t; w)$ of the corresponding WIP-level. For the various experiments in this paper, we assume that the EPT distributions for each WIP level are gamma distributed, with mean EPT $t_{\mathrm{e}}(w)$ and coefficient of variation of the EPT $c_{\mathrm{e}}(w)$. We choose the gamma distribution because it fits the empirical data well in the considered cases, and it is characterized by two parameters only, being its mean $t_{\mathrm{e}}(w)$ and its coefficient of variation $c_{\mathrm{e}}(w)$. It is advised to practitioners to validate which distribution is the most suitable for the workstation being modeled.

Overtaking realizations are also grouped, but now according to the number of lots in the system w *upon arrival*. In this case, we do not define a maximum WIP-level. For each WIP-level, we again obtain a distribution which is used for the overtaking distribution $F_{\mathrm{K}}(k; w)$.

D. Implementation Issues

To obtain arrival and departure events for an operational workstation, a procedure similar as described in [21] is used. From the data storage system in the fab, typically the manufacturing execution system (MES), the status history of lots processed during some time period is obtained. Most lots that arrive at a workstation first have to wait in a buffer, and are subsequently processed on one of the machines in the workstation. For these "regular" lots, we define an arrival as the start of the waiting period, and the departure as the end of the processing period.

However, exceptions to this common situation may occur. One exception occurs when a lot temporarily gets the status "on hold" while it is in the buffer; the "on hold" status means that the lot is unavailable for processing because of a quality problem. For such a hold lot, we define the lot arrival to occur after the "on hold" status has finished, and the lot starts waiting uninterruptedly for processing. As for normal lots, a departure is defined as the time the lot departs from the workstation. Another exception is merging of lots. Wafers arrive in different front opening unified pods (FOUPs) but are (re)united into one FOUP and processed together. In this case, the arrival of the (re)united lot is defined to occur when the last set of wafers arrives. The departure occurs when the reunited lot has finished processing.

In semiconductor practice only a limited number of arrival and departure events may be available. MES data may be stored only for a couple of weeks, or structural changes to the workstation occurred (e.g., an additional machine was installed), which makes data of only a few weeks representative for the workstation's behavior. As a consequence, it is more difficult to accurately estimate the mean EPT $t_{\mathrm{e}}(w)$, the coefficient of variation of the EPT $c_{\mathrm{e}}(w)$, and $F_{\mathrm{K}}(k; w)$. Consequently, the cycle time predictions may deteriorate. In particular, we observe that an accurate estimate of t_{e} for maximum WIP-level $w_{\max}$ is crucial. The reason is that $1/t_{\mathrm{e}}(w_{\max})$ determines the predicted maximum throughput of the workstation. To arrive at an accurate $t_{\mathrm{e}}(w_{\max})$ estimate, we take for $w_{\max}$ the WIP-level above which $t_{\mathrm{e}}(w)$ is approximately constant. If we set $w_{\max}$ to this WIP-level, we obtain the largest number of EPT realizations for $w_{\max}$, while we do not discard the WIP-dependency of t_{e}.

Also for WIP-levels smaller than $w_{\max}$ we observe noise in $t_{e(w)}$ and $c_{e(w)}$ due to the small number of EPT realizations that may have been collected for certain WIP-levels. It may even occur that for some WIP levels, no EPT realizations are obtained at all. To overcome this difficulty, a curve fitting approach similar to [21] is introduced. We approximate the measured $t_e(w)$ values by $\hat{t}_e(w)$, for which we use the following exponential function [21]:

$$\hat{t}_e(w) = \theta + (\eta - \theta)e^{-\lambda(w-1)} \tag{1}$$

where θ represents the value of $\hat{t}_e(w)$ at $w = \infty$. Variable η represents the value of $\hat{t}_e(w)$ at $w = 1$. Variable λ represents the "decay constant" of the exponential curve. We set θ equal to the measured t_e for $w = w_{\max}$. Variables η and λ are estimated using a nonlinear least-squares fitting procedure, in which the $t_e(w)$ estimates are weighted according to $\sqrt{n(w)}$, with $n(w)$ the number of EPT realizations obtained for WIP level w.

Similarly, we approximate $c_e(w)$ by $\hat{c}_e(w)$ for which we also use exponential function (1). For the overtaking probabilities, we do not introduce a curve fit, but use the measured overtaking probabilities directly in the aggregate model. For WIP levels lower than the lowest WIP level for which we obtained overtaking probabilities, we assume that no overtaking occurs. For higher WIP levels, we use the same overtaking probabilities as measured for the highest WIP level.

In principle, curve fitting is also desirable to represent the overtaking probabilities. In [22], discrete distributions are fitted for which the stochastic variable has values in the range $[0, 1, \ldots, \infty]$. However, in our case a sampled K value is always less than or equal to a finite value (w). For this particular type of distribution, very few results on distribution fitting procedures are available.

III. Validation

Two simulation test cases are presented to validate the proposed method. The first case is used to investigate the accuracy of the method in predicting cycle time distributions for various workstation configurations. In this case, it is assumed that sufficient measured arrivals and departures are available to accurately estimate the aggregate model parameters. The second case is used to investigate the predictions for a workstation representing a lithography workstation that produces two different product types, and for which a limited amount of measured arrival and departure events is available. The two simulation cases, and the aggregate model used in this section are implemented as a discrete-event simulation model in the language χ [23].

A. Case I

1) *Description:* Case I is depicted in Fig. 1(a). The workstation consists of m identical parallel machines. Each machine consists of l sequential integrated processes so may be viewed as a cascading machine. Each integrated process has a gamma-distributed process time with mean t_0 and coefficient of variation c_0. Lots arrive at the infinite buffer preceding the workstation; the interarrival times are independent and follow a gamma distribution with mean t_a and coefficient of variation c_a. The order in which lots in the buffer are processed is defined by dispatching rule d. If more than one machine is available for processing, the lot is sent to the machine of which the first process has the longest idle time (fairness).

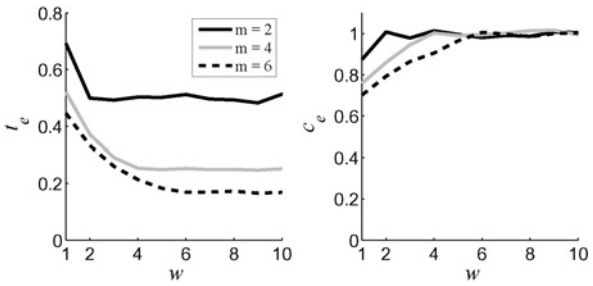

Fig. 3. Mean EPT t_e and CV c_e as a function of WIP level w for case I for different values of m, and constant $l = 1$, $c_0 = c_a = 1.0$, and d = FCFS.

We experiment with different values of m, l, c_0, and c_a. For the dispatching rule d, we consider FCFS, non-preemptive last-come-first-served (LCFS), and priority (Pr) dispatching. For FCFS and LCFS dispatching, we assume that all lots have the same mean process time $t_0 = 1.0$, and coefficient of variation of the process time c_0 in the various processes. For Pr dispatching, we use two lot classes. Class A requires $t_0 = 1.0$, whereas class B requires $t_0 = 2.0$. Coefficient of variability c_0 is again the same for all lots. Class A has non-preemptive priority over class B.

2) *Estimating Model Parameters:* To estimate the WIP-dependent EPT distribution $F_E(t; w)$ and overtaking distribution $F_K(k; w)$ for a workstation configuration, we obtained arrivals and departures of 10^6 lots at a throughput ratio $\delta/\delta_{\max}$ of 0.8, with $\delta = 1/t_a$ the throughput of the workstation and $\delta_{\max}$ the maximum obtainable throughput of the workstation.

The algorithm in Appendix A is used to calculate EPT realizations, which are grouped according to WIP-levels, as explained in Section II-C. Recall that the EPT distribution for each WIP-level is represented by a gamma distribution with mean t_e and coefficient of variation c_e. Distribution parameters t_e and c_e are obtained directly from the measured data. Also recall that maximum WIP-level $w_{\max}$ groups all EPTs that started with WIP-level $w \geq w_{\max}$. In this simulation case, we use an automated procedure to determine $w_{\max}$; we choose $w_{\max}$ as high as possible, under the condition that the half-width of the 95% confidence interval of $t_{e,w_{\max}}$ is less than 1% of the sample mean.

The algorithm in Appendix A also yields overtaking realizations k, which are grouped according to WIP-levels as well. For each WIP-level, we use the empirical overtaking distribution directly in the aggregate model so we do not use a distribution fit here (see Section II-D).

To illustrate the proposed method, we now present the measured EPT-distribution parameters and the measured overtaking probabilities for a selection of workstation configurations. Fig. 3 shows mean EPT t_e (left-hand side) and coefficient of variation of the EPT c_e (right-hand side) as a function of the WIP w for $m = 2$, 4, and 6, with $l = 1$, $c_0 = c_a = 1.0$, and d = FCFS. Mean EPT t_e decreases for increasing w,

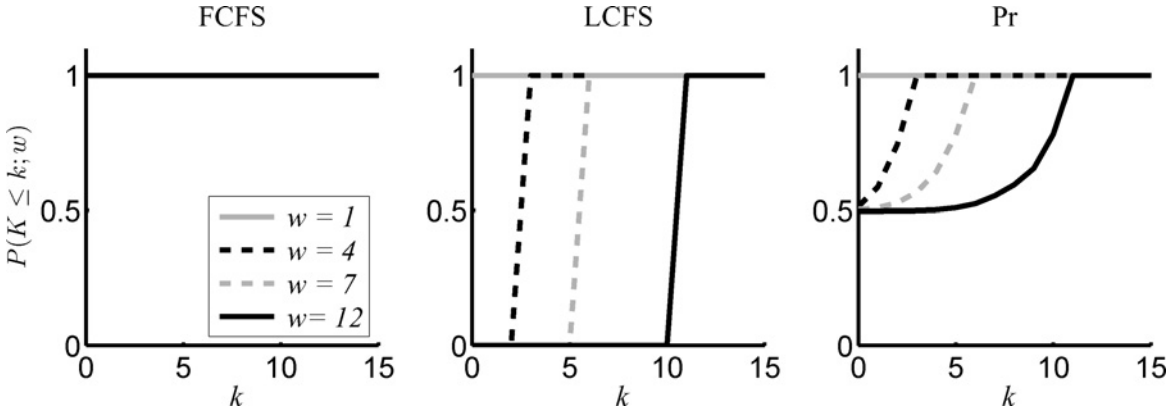

Fig. 4. Case I: cumulative probability for a newly arrived lot to overtake K lots already in the system for various WIP-levels w and for different dispatching rules, with $m = l = 1$, and $c_0 = c_a = 1.0$.

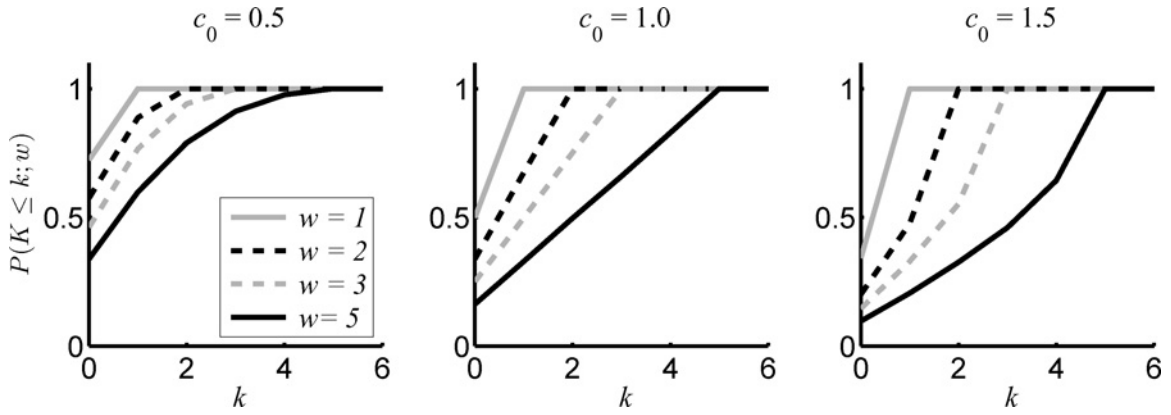

Fig. 5. Case I: cumulative probability for a newly arrived lot to overtake K lots already in the system for various WIP-levels w and for different values of arrival coefficient of variability c_0, with $m = 6$, $l = 1$, $c_a = 1.0$, and d = FCFS.

until $w \approx m$. For $w > 1$ the mean EPT may be interpreted as the mean interdeparture time of lots at the workstation. For increasing w, more parallel machines are processing, up to the maximum number of machines m. Hence, the mean interdeparture time decreases up to $w = m$.

Coefficient of variation c_e increases until $w \approx m$, where c_e reaches the value 1.0 corresponding to an exponential interdeparture time. Intuitively, the standard deviation of the EPT σ_e is expected to decrease. The behavior of c_e, however, is less obvious, since it depends on both t_e and σ_e. Apparently, in this particular case, c_e increases because t_e decreases faster than σ_e, but we also observed that in other examples c_e may even exhibit non-monotonous behavior.

Next, we show that the overtaking distribution $F_K(k; w)$ depends on the dispatching rule. Fig. 4 shows the cumulative overtaking probabilities $P(K \leq k; w)$ as a function of k for several values of WIP-level w. We consider FCFS, LCFS, and Pr dispatching with $m = l = 1$, and $c_0 = c_a = 1.0$. For $m = 1$, overtaking only occurs due to the dispatching rule and not due to parallel processing. In the FCFS case (the left-hand plot) $P(K \leq k; w) = 1$ for all values of k and w, so lots do not overtake. In the (non-preemptive) LCFS case (the middle plot), $P(K \leq k; w)$ jumps from 0 to 1 for $k = w-1$, so each arriving lot overtakes all lots in the system, except the one in process. For Pr dispatching (the right-hand plot), the probability to overtake no lots is 0.5 for $w > 1$, because 50% of the arriving lots is of type B (with long process times), which do not overtake. The type A lots may overtake one or more type B lots in the buffer, with a maximum of the total amount of lots in the system, minus the lot in process. Therefore, the cumulative probability reaches 1.0 for $k = w - 1$.

Fig. 5 shows that the overtaking probabilities depend on c_0. In Fig. 5 we consider $c_0 = \{0.5, 1.0, 1.5\}$, with $m = 6$, $l = 1$, $c_a = 1.0$, and d = FCFS. For this configuration, overtaking only takes place due to parallel processing. Hence, in all three plots of Fig. 5 the maximum number of lots that can be overtaken is 5. For $c_0 = 1.0$ (the middle plot), there is an equal probability to overtake $K = 0, ..., \min(w, 5)$ lots already in the system due to the exponential process times, which makes the cumulative probability to increase linearly. For $c_0 = 0.5$ (the left-hand plot), the slope of the cumulative overtaking probability curve decreases for increasing k, indicating that the overtaking probability decreases for increasing k. This is because the process time variability is low compared to the case in which $c_0 = 1.0$, so less overtaking occurs. For $c_0 = 1.5$ (the right-hand plot), the slope of the curves increases for increasing k, because the servers have a relatively high process time variability, so more overtaking occurs.

3) *Cycle Time Predictions:* The detailed simulation model of the considered workstation is used to measure the "real" cycle time distribution for various workstation configurations for throughput ratios δ/δ_{max} ranging from 0.3 to 0.95. For each

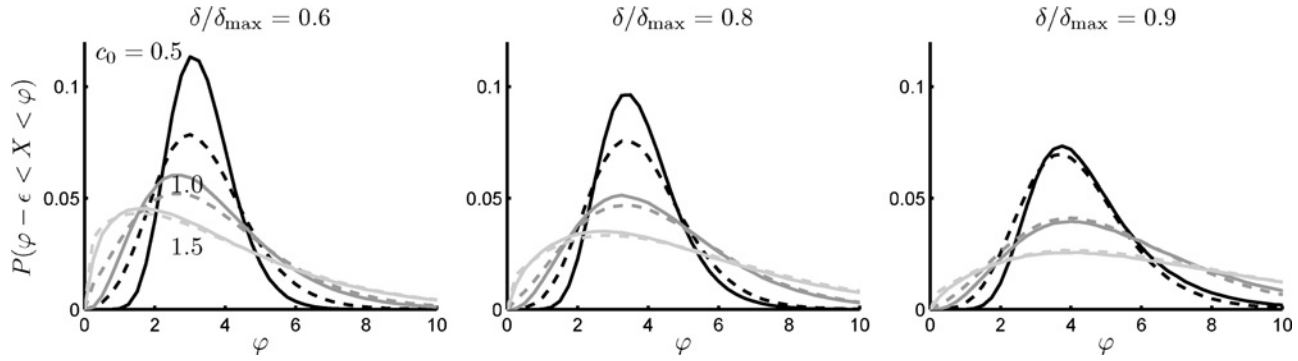

Fig. 6. Cycle time distribution of the workstation (the solid curves), and predicted by the aggregate model (the dashed curves) for $c_0 = \{0.5, 1.0, 1.5\}$, with $m = 5$, d = FCFS, $l = 3$, and $c_a = 1.0$. The black curves consider $c_0 = 0.5$, the dark grey curves consider $c_0 = 1.0$, and the light grey curves represent $c_0 = 1.5$.

throughput ratio, 30 simulation replications of 10^5 processed lots are performed. For each replication run, the first $2 \cdot 10^4$ lots are discarded to account for the start-up phenomenon.

For each considered workstation configuration, we use the aggregate model depicted in Fig. 1(b) to predict cycle time distributions. The aggregate model is trained at $\delta/\delta_{\max} = 0.8$ using 10^6 arrivals and departures measured at the detailed workstation model. We predict the cycle time distribution for the same throughput levels for which we calculated the real cycle time distribution, using again 30 replications, a simulation length of 10^5 lots, and a start-up period of $2 \cdot 10^4$ lots. For the arrival process in the aggregate model we use a gamma distribution with mean t_a depending on the considered throughput level. For the coefficient of variation c_a we choose the same value as in the workstation. In the aggregate model we use gamma EPT distributions for each WIP level, of which the shape and scale parameters are determined from the measured t_e and c_e values for the corresponding WIP levels w. For the overtaking distributions in the aggregate model, we directly use the empirical overtaking distribution. We measure the empirical overtaking distribution for WIP-levels up to a certain value. For higher WIP-levels, we assume in the aggregate model that the overtaking probabilities are the same as for the highest measured WIP-level.

First, we investigate the influence of the process time variability (c_0) on the prediction accuracy of the cycle time distribution by the aggregate model. The results are depicted in Fig. 6 and Table I. Fig. 6 depicts cycle time distributions of the workstation (the solid curves), and cycle time distributions predicted by the aggregate model (the dashed curves) for workstation configurations with $c_0 = \{0.5, 1.0, 1.5\}$, with $m = 5$, $l = 3$, d = FCFS, and $c_a = 1.0$. We do not show the confidence intervals on the cycle time distributions because they are very small. From left to right the figure shows distributions for throughput ratios of 0.6, 0.8, and 0.9, respectively. Recall that $\delta/\delta_{\max} = 0.8$ is the training level. The different curves in the plots correspond to different values of c_0: the top solid and dashed black curves in each plot correspond to $c_0 = 0.5$, the middle dark grey curves correspond to $c_0 = 1.0$, and the bottom light grey curves correspond to $c_0 = 1.5$. The x-axis denotes the cycle time φ, whereas the y-axis denotes the probability $P(\varphi - \epsilon < X < \varphi)$, where ϵ denotes the size of an interval, for which we choose 0.25.

TABLE I

MEAN AND 95% QUANTILE OF THE CYCLE TIME DISTRIBUTION MEASURED AT THE CONSIDERED WORKSTATION, AND PREDICTED BY THE PROPOSED AGGREGATE MODEL FOR $c_0 = \{0.5, 1.0, 1.5\}$, WITH $m = 5$, d = FCFS, $l = 3$, AND $c_a = 1.0$

	Mean					
	$c_0 = 0.5$		$c_0 = 1.0$		$c_0 = 1.5$	
$\delta/\delta_{\max}$	Meas.	Pred.	Meas.	Pred.	Meas.	Pred.
0.3	3.02	2.92	3.08	2.82	3.18	2.61
0.6	3.21	3.25	3.48	3.51	3.80	3.77
0.7	3.37	3.41	3.74	3.79	4.18	4.23
0.8	3.66	3.65	4.21	4.21	4.90	4.95
0.85	3.93	3.84	4.65	4.54	5.59	5.54
0.9	4.47	4.08	5.50	5.02	7.02	6.51
0.95	6.08	4.47	7.90	5.81	11.48	8.52
	95% Quantile					
	$c_0 = 0.5$		$c_0 = 1.0$		$c_0 = 1.5$	
$\delta/\delta_{\max}$	Meas.	Pred.	Meas.	Pred.	Meas.	Pred.
0.3	4.57	5.68	6.40	6.63	8.48	7.18
0.6	4.84	5.67	6.98	7.45	9.51	9.41
0.7	5.10	5.81	7.41	7.79	10.17	10.17
0.8	5.69	6.12	8.29	8.45	11.54	11.44
0.85	6.33	6.44	9.21	9.04	13.06	12.62
0.9	7.76	6.92	11.26	9.97	16.60	14.73
0.95	12.47	7.80	17.71	11.71	29.44	19.67

Table I presents the mean and the 95% quantile of the cycle time distribution of the workstation, and the mean and quantile predicted by the aggregate model for varying process time variability (c_0). Results are given for throughput ratios $\delta/\delta_{\max}$ from 0.3 to 0.95. The half-width of the confidence interval of the values in the table depends on the throughput ratio; for $\delta/\delta_{\max} \leq 0.90$ the half-widths of the confidence intervals are smaller than 2.5% of the sample mean for all experiments. For $\delta/\delta_{\max} = 0.95$, the confidence intervals are smaller than 6.5%.

Fig. 6 shows that for $c_0 = 1.0$ and $c_0 = 1.5$, the predicted cycle time distributions are close to the cycle time distributions measured at the workstation being modeled, for all considered throughput levels. For $c_0 = 0.5$, the accuracy of the predicted cycle time distribution deteriorates for decreasing throughput ratio, in particular for relatively short cycle times. The measured cycle time distribution shows less variability than the predicted cycle time distribution. The reason may be that the EPT and the number of overtaken lots in the aggregate model are sampled independently for successive lots, which possibly creates more variability than occurs in reality.

TABLE II
MEAN AND 95% QUANTILE OF THE CYCLE TIME DISTRIBUTION MEASURED AT THE CONSIDERED WORKSTATION, AND PREDICTED BY THE PROPOSED AGGREGATE MODEL FOR $m = \{1, 3, 5\}$, WITH d = FCFS, $l = 3$, d = FCFS, AND $c_a = 1.0$

			Mean			
	$m = 1$		$m = 3$		$m = 5$	
δ/δ_{max}	Meas.	Pred.	Meas.	Pred.	Meas.	Pred.
0.3	3.66	3.45	3.14	2.81	3.08	2.82
0.6	5.12	5.26	3.69	3.76	3.48	3.51
0.7	6.19	6.38	4.07	4.19	3.74	3.79
0.8	8.28	8.30	4.81	4.84	4.21	4.21
0.85	10.36	9.86	5.51	5.38	4.65	4.54
0.9	14.47	12.35	6.87	6.26	5.50	5.02
0.95	26.33	16.61	10.83	7.91	7.90	5.81
			95% Quantile			
	$m = 1$		$m = 3$		$m = 5$	
δ/δ_{max}	Meas.	Pred.	Meas.	Pred.	Meas.	Pred.
0.3	7.77	8.30	6.52	6.74	6.40	6.63
0.6	11.50	12.56	7.44	8.11	6.98	7.45
0.7	14.53	15.47	8.19	8.83	7.41	7.79
0.8	20.64	20.75	9.90	10.08	8.29	8.45
0.85	26.72	25.14	11.65	11.36	9.21	9.04
0.9	38.95	32.10	15.48	13.67	11.26	9.97
0.95	74.44	43.50	26.94	18.61	17.71	11.71

TABLE III
MEAN AND 95% QUANTILE OF THE CYCLE TIME DISTRIBUTION MEASURED AT THE CONSIDERED WORKSTATION, AND PREDICTED BY THE PROPOSED AGGREGATE MODEL FOR $d = \{$FCFS, LCFS, Pr$\}$, WITH $m = 3$, $l = 3$, $c_0 = 1.0$, and $c_a = 1.0$

			Mean			
	d = FCFS		d = LCFS		d = Pr	
δ/δ_{max}	Meas.	Pred.	Meas.	Pred.	Meas.	Pred.
0.3	3.15	2.81	3.15	2.82	4.77	4.19
0.6	3.70	3.77	3.70	3.78	5.69	5.82
0.7	4.09	4.21	4.09	4.21	6.29	6.51
0.8	4.85	4.88	4.85	4.89	7.40	7.43
0.85	5.58	5.42	5.58	5.42	8.39	8.04
0.9	7.05	6.28	7.04	6.24	10.34	8.82
0.95	11.52	7.84	11.47	7.65	15.95	9.97
			95% Quantile			
	d = FCFS		d = LCFS		d = Pr	
δ/δ_{max}	Meas.	Pred.	Meas.	Pred.	Meas.	Pred.
0.3	6.52	6.74	6.52	6.75	11.13	10.56
0.6	7.45	8.13	7.50	8.26	12.75	13.37
0.7	8.25	8.86	8.29	9.13	14.15	14.72
0.8	9.99	10.20	9.90	10.89	17.48	16.93
0.85	11.87	11.40	11.68	12.56	21.06	18.60
0.9	15.96	13.65	15.68	15.32	29.38	21.01
0.95	28.77	17.97	25.84	19.92	54.61	25.13

Table I confirms the observations obtained from Fig. 6: for $c_0 = 1.0$ and $c_0 = 1.5$ the mean and 95% quantile are predicted with reasonable accuracy for throughput ratios from 0.6 to 0.9. For $c_0 = 0.5$, the prediction accuracy of the 95% quantile deteriorates. Note, however, that this dependency of the prediction accuracy on c_0 seems not to be present for the mean cycle time. Furthermore, Table I indicates that the predictions of the mean and 95% quantile become inaccurate for $\delta/\delta_{max} = 0.95$. For this throughput ratio, the prediction is very sensitive to the measured EPT and overtaking distribution. There is clearly a range of throughput levels around the training point where accurate predictions are obtained; further away from the training point, the accuracy deteriorates.

Next, we investigate the influence of the number of parallel machines in the workstation on the prediction accuracy. Table II presents the means and 95% quantiles of the cycle time distribution of the workstation, and the means and 95% quantiles predicted by the aggregate model. The half-widths of the confidence intervals of the values in the table are similar to the those of the values in Table I. We consider $m = \{1, 3, 5\}$, with $l = 3$, d = FCFS, and $c_0 = c_a = 1.0$. Table II shows that for throughput ratios $\delta/\delta_{max} > 0.8$, the prediction errors of the mean and 95% quantile in case $m > 1$ are less than in case $m = 1$. So for this type of workstation, the throughput range for which accurate predictions can be made is larger for a multi-machine workstation than for a single machine workstation.

Table III visualizes the effect of various dispatching rules on the prediction accuracy of the mean and 95% quantile. We experimented with $d = \{$FCFS, LCFS, Pr$\}$, with constant $m = 3$, $l = 3$, and $c_0 = c_a = 1.0$. Again, the half-widths of the confidence intervals of the values in the table are similar to those of the values in Table I. The table shows that the prediction errors of the mean are similar for all three dispatching rules. However, for $\delta/\delta_{max} \geq 0.85$, the prediction accuracy of the 95% quantile is less accurate for the Pr dispatching rule than for the FCFS and LCFS dispatching rules. The reason is that for the workstation with Pr dispatching, the number of lots overtaken by an arriving lot depends on its class, and on the classes of the lots that are already present in the workstation. These dependencies are not taken into account in the aggregate model.

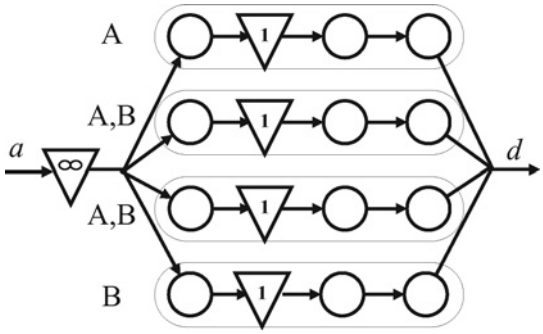

Fig. 7. Case II representing a lithography workstation.

We have also experimented with different numbers of integrated processes l, being $l = 2$, and $l = 4$, and different values of the coefficient of variability of the interarrival times c_a, being 0.5 and 1.5. We observe that l and c_0 have little influence on the accuracy of the cycle time predictions. The value of c_a has little influence, because we also use c_a for the arrival process in the aggregate model.

B. Case II

1) *Description:* Case II is depicted in Fig. 7. The setup of Case II may be viewed as a group of track-scanner lithography tools. 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 FCFS order taking into account machine recipe qualification. The first machine is qualified only for recipe A, the second and third

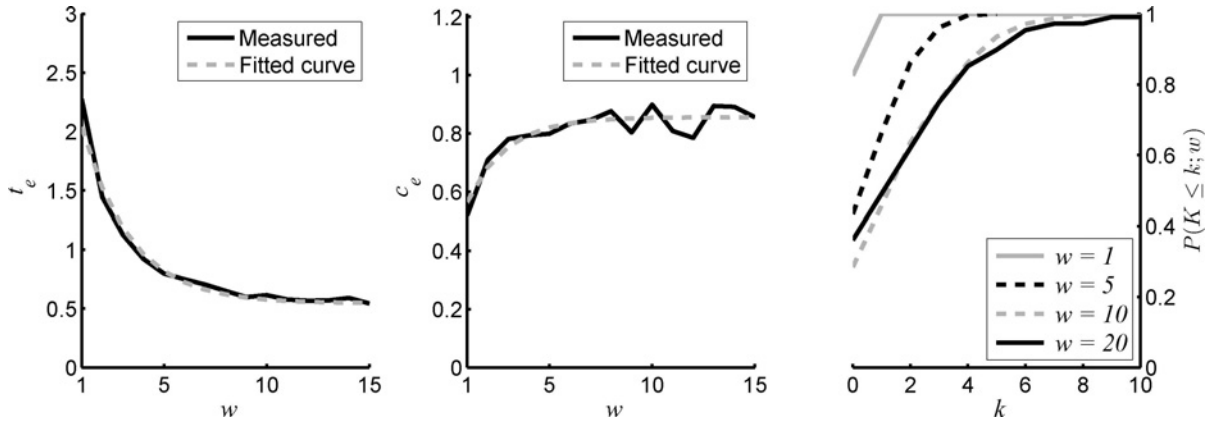

Fig. 8. Measured and fitted mean EPT t_e (left), coefficient of variability c_e (middle), and cumulative overtaking probabilities (right) for case II using 10^4 arrivals and departures.

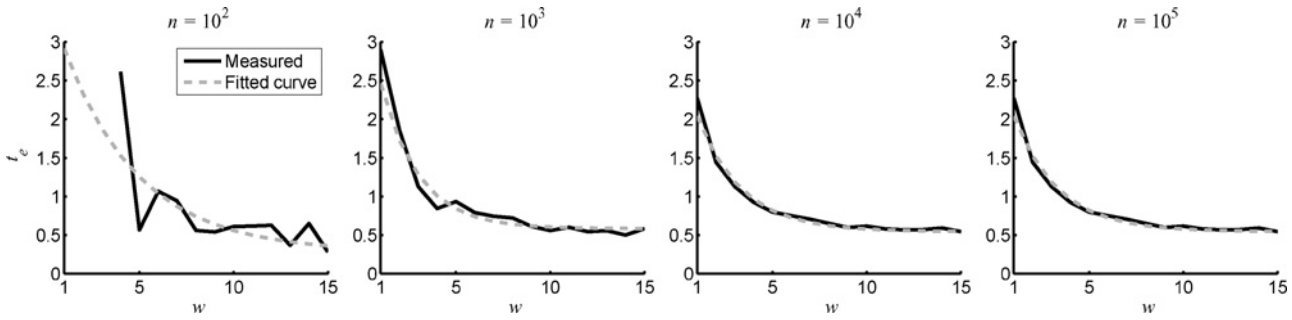

Fig. 9. Measured and fitted mean EPT t_e for case II using 10^2, 10^3, and 10^4 arrivals and departures.

machine are qualified for recipe A and B, and the fourth machine is qualified only for recipe B. 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 can be viewed as the track and are assumed to have a constant process time of 1.0. The second process step may be viewed as the scanner and is assumed to have an exponential process time distribution with mean 2.0.

Unlike case I, we now measure arrivals and departures of far less than 10^6 lots, because this amount of lots is typically not available in semiconductor manufacturing practice. We denote the number of measured arrival and departure events by n; we experiment with different values of n, being $n = \{10^2, 10^3, 10^4, 10^5\}$, respectively.

2) *Calculating Model Parameters:* For each value of n, arrival and departure events were obtained at a throughput ratio of $\delta/\delta_{\max} = 0.8$. We again use the algorithm given in Appendix A to calculate EPT realizations and overtaking realizations K, which were grouped according to WIP-levels, as explained in Section II-B. We also use gamma distributions to represent the EPT distributions for each WIP-level.

Fig. 8 shows the estimated model parameters using $n = 10^4$ measured arrivals and departures. The left plot in Fig. 8 shows $t_e(w)$ (the black curve). The middle plot of Fig. 8 shows $c_e(w)$ (the black curve). We choose $w_{\max} = 15$, because for $w > 15$, $t_e(w)$ is approximately constant. The dashed grey curves in the left and middle plots represent the fitted curves $\hat{t}_e(w)$ and $\hat{c}_e(w)$, respectively, which are calculated using the curve fitting procedure explained in Section II. For $\hat{t}_e(w)$, the values of curve fit parameters θ, η, and λ become 2.224, 0.548, and 0.4716, respectively. For $\hat{c}_e(w)$, θ, η, and λ become 2.224, 0.548, and 0.4716, respectively.

The right plot of Fig. 8 shows the cumulative overtaking probabilities $P(K \leq k; w)$ as a function of k for several values of w, using $n = 10^4$. We do not introduce a curve fit; we use the measured overtaking distribution directly in the aggregate model. For WIP levels lower than the WIP levels for which we measured the overtaking probabilities, we again assume that no overtaking occurs; for higher WIP levels we assume that the overtaking probabilities are the same as for the highest measured WIP-level.

Fig. 9 depicts $t_e(w)$ and $\hat{t}_e(w)$ estimated using, from left to right, $n = 10^2$, $n = 10^3$, and $n = 10^4$. The figure shows that for a decreasing amount of measured events, the noise in the values of $t_e(w)$ increases. Also, for the lowest $n = 10^2$ no EPT estimates were obtained at low WIP levels. The fitted curve smooths the noise and provides estimates for the mean EPT at the low WIP levels at $n = 10^2$ for which no $t_e(w)$ estimates were measured (by means of extrapolation). The figure shows that for $n = 10^2$, the mean EPT estimated by the fitted curve at $w_{\max} = 15$, $\hat{t}_e(15)$, is considerably lower than $\hat{t}_e(15)$ for $n = 10^3$ and $n = 10^4$. The cause is that for decreasing n, it becomes increasingly difficult to accurately estimate the mean EPT at $w_{\max} = 15$, because few EPT realizations are obtained for $w_{\max}$.

For $c_e(w)$ and $\hat{c}_e(w)$, the amount of noise increases for decreasing n as well, and $c_e(w)$ are also missing for low WIP levels in case $n = 10^2$.

3) *Cycle Time Predictions:* The detailed simulation model of the case II workstation is used to calculate the workstation's cycle time distribution for throughput ratios δ/δ_{max} ranging from 0.3 to 0.95. Recall that the training level is $\delta/\delta_{max} = 0.8$. We use the same number of replications, simulation length, and start-up period as in case I.

The aggregate model depicted in Fig. 1(b) is used to predict cycle time distributions, using $n = 10^2$, $n = 10^3$, $n = 10^4$, and $n = 10^5$ measured arrival and departure events, respectively, to estimate the aggregate model parameters. We again use the same number of replications, simulation length, and start-up period as in case I. In the aggregate model, we use Poisson arrivals as we did in the detailed simulation model, but assume all lots are the same (no recipes are used). We use gamma EPT distributions in the aggregate model for each WIP level w, with the fitted mean $\hat{t}_e(w)$ and coefficient of variability $\hat{c}_e(w)$. For the overtaking distributions in the aggregate model, we use the empirical overtaking distributions (as we did in case I).

Table IV presents the mean and 95% quantile of the workstation, and the mean and 95% quantile predicted by the aggregate model for $n = \{10^2, 10^3, 10^4, 10^5\}$ for throughput ratios ranging between 0.3 and 0.95. For $\delta/\delta_{max} \leq 0.90$ the half-widths of the confidence intervals are typically smaller than 2.5% of the sample mean for all experiments. For $\delta/\delta_{max} = 0.95$, the confidence intervals are smaller than 5.5%. The horizontal bars for $\delta/\delta_{max} = 0.95$ and $n = 10^3$ indicate that the aggregate model simulation was instable (the arrival rate is higher than the maximum processing rate). The reason is that when relatively few EPT realizations are obtained, the curve fit may overestimate the maximum capacity of the system. The table shows that at the training level ($\delta/\delta_{max} = 0.80$), the prediction accuracy of the mean and 95% quantile seems to be independent of n. Even if n is only 10^2, the mean and 95% quantile can still be predicted within 10% accuracy at the training level. However, for $n = 10^2$, the throughput range for which the mean and 95% quantile are predicted accurately is very small. In particular, the accuracy of predictions higher than the training level benefits from increasing n. The fact that the workstation processes two different product types does not seem to influence the results.

In the various experiments performed in this section, we observed that the calculation time required to evaluate the aggregate simulation model is about 20 times shorter than the calculation time required to evaluate the detailed simulation model.

TABLE IV
MEAN AND 95% QUANTILE OF THE CYCLE TIME DISTRIBUTION OF THE CASE II WORKSTATION, AND PREDICTED BY THE PROPOSED AGGREGATE MODEL FOR $n = \{10^2, 10^3, 10^4, 10^5\}$

			Mean					
	$n = 10^2$		$n = 10^3$		$n = 10^4$		$n = 10^5$	
δ/δ_{max}	Meas.	Pred.	Meas.	Pred.	Meas.	Pred.	Meas.	Pred.
0.3	4.26	6.43	4.26	4.46	4.26	3.90	4.26	3.80
0.6	5.13	6.60	5.13	5.33	5.13	5.04	5.13	5.01
0.7	5.67	6.43	5.67	5.86	5.67	5.48	5.67	5.48
0.8	6.60	6.28	6.60	7.13	6.60	6.27	6.60	6.35
0.85	7.48	6.22	7.48	8.80	7.48	7.02	7.48	7.25
0.9	9.15	6.16	9.15	14.25	9.15	8.35	9.15	8.99
0.95	14.00	6.12	14.00	–	14.00	12.31	14.00	14.56
			95% Quantile					
	$n = 10^2$		$n = 10^3$		$n = 10^4$		$n = 10^5$	
δ/δ_{max}	Meas.	Pred.	Meas.	Pred.	Meas.	Pred.	Meas.	Pred.
0.3	8.72	15.36	8.72	9.50	8.72	8.68	8.72	8.59
0.6	10.80	13.89	10.80	10.76	10.80	10.32	10.80	10.32
0.7	11.91	13.14	11.91	11.92	11.91	11.14	11.91	11.22
0.8	13.88	12.63	13.88	14.95	13.88	12.85	13.88	13.08
0.85	15.89	12.41	15.89	19.41	15.89	14.64	15.89	15.28
0.9	20.04	12.18	20.04	35.14	20.04	17.88	20.04	19.88
0.95	33.35	11.95	33.35	–	33.35	29.71	33.35	35.94

IV. CROLLES2 CASE

We finally apply the proposed method to an operational workstation at the Crolles2 wafer fab. Crolles2 is a multi-product 300 mm fab in which both high volume products and small series and prototype products are produced. The production lots are called FOUPs and can contain up to 25 wafers. In the data collection period, approximately 80% of the FOUPs contained the maximum of 25 wafers; the other 20% of the FOUPs contained less than 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, from the arrival and departure data we calculate the EPT distributions and overtaking probability distributions. Finally, cycle time distributions are predicted using the aggregate model, which is implemented as a discrete-event simulation model in the language χ [23].

A. Crolles2 Lithography Workstation

The lithography workstation consists of 14 track-scanner machines of different types, with different recipe qualifications. Lots are loaded onto one of the load ports of a machine, whereupon 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.

B. Calculating Model Parameters

At the Crolles2 site, arrivals and departures of 42 141 lots processed at the litho workstation were obtained from the MES. The MES data is filtered as described in Section II. After this filtering, the EPT algorithm in Appendix A is used to calculate EPT realizations and lot overtaking realizations. We choose $w_{max} = 100$, because for $w > 100$, $t_e(w)$ does not decrease further. Similar to Section III, we use the gamma distribution to represent the EPT distributions for each WIP-level.

The left plot of Fig. 10 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 II using exponential function

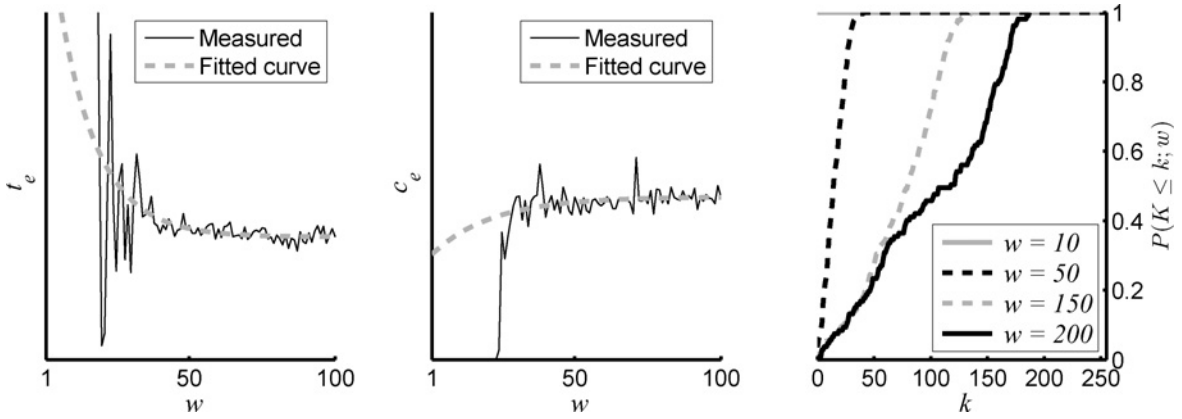

Fig. 10. Measured and fitted mean EPT t_e (left), coefficient of variability c_e (middle), and cumulative overtaking probabilities (right) of the Crolles2 lithography workstation.

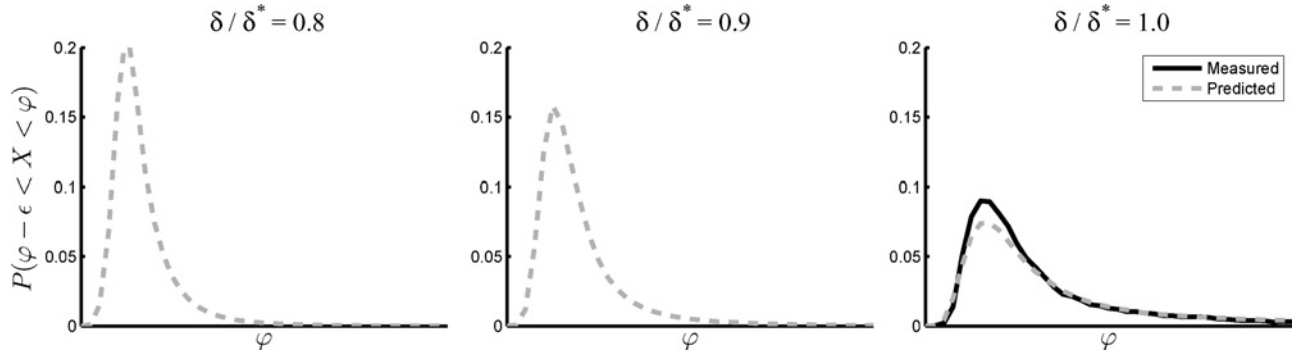

Fig. 11. Measured and predicted cycle time distribution of the litho workstation, for relative throughput levels of 0.8, 0.9, and 1.0.

(1). Note that we do not have EPT realizations for $w < 18$; the t_e and c_e estimates for these WIP levels are estimated by the fitted curve by extrapolation.

The left plot of Fig. 10 clearly illustrates that the mean interdeparture time decreases as w increases: 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 Fig. 10 shows the measured cumulative overtaking probabilities $P(K \leq k; w)$. Note that for $w \geq 50$ considerable overtaking occurs. We have not measured overtaking realizations for WIP-levels either lower than 18 or higher than 256. We assume that no overtaking takes place for WIP-levels lower than 18. For WIP levels higher than 256, we use the same overtaking probabilities as measured for a WIP-level of 256.

C. Cycle Time Predictions

We use the aggregate model depicted in Fig. 1(b) to estimate cycle time distributions of the lithography workstation, using gamma-distributed EPT distributions based on fitted values $\hat{t}_e(w)$ and $\hat{c}_e(w)$, and the empirical overtaking distribution as model parameters. We again perform 30 simulation replications, a simulation run length of 10^5 lots, a start-up period of $2 \cdot 10^4$ lots, and the same arrival process as measured at the lithography workstation.

Fig. 11 depicts cycle time distributions for the lithography workstation at relative throughput levels 0.8, 0.9, and 1.0. The relative throughput is defined here as the throughput δ divided by the throughput at the training point δ^*. We use the relative throughput instead of throughput ratio $\delta/\delta_{\max}$ for confidentiality reasons. We do not consider relative throughput levels higher than 1.0, because δ^* is already very high.

The rightmost plot represents the cycle time distribution at the training point of the workstation ($\delta/\delta^* = 1$). The x-axis denotes cycle time φ, the y-axis probability $P(\varphi - \epsilon < X < \varphi)$ (for some small $\epsilon > 0$). The solid line in the rightmost plot represents the measured cycle time distribution of the workstation at the training point. The dashed lines represent the cycle time distributions estimated by the proposed method.

Fig. 11 shows that in particular the tail of the cycle time distribution is accurately estimated at the training point (the rightmost plot). For a decreasing relative throughput level, the accuracy of the predicted cycle times decreases. We can only verify the cycle time distribution at the training point. The simulation test cases described in Section III indicate that accurate predictions can be made for throughput levels other than the training point, in particular, for the tails of the distributions. Therefore, we expect that accurate cycle time distributions can be obtained for throughput levels in a range around the training level.

D. Re-Estimation of Model Parameters

To effectively implement the EPT-based aggregate modeling method in practice, the EPT and overtaking distribution need to be occasionally re-estimated to account for the changing

conditions in the workstation. In case the throughput of the workstation changes, the EPT and overtaking distributions do not have to be re-estimated. Throughput fluctuations actually improve the aggregate model parameter estimates, because EPT and overtaking realizations can be obtained for a larger WIP range. In case conditions temporarily change (i.e., during a period much shorter than the data collection period), such as a machine that is down or in maintenance, the EPT and overtaking distribution do not have to be re-estimated either. These changes may be viewed as "stochastic" behavior of the workstation. However, for long-term, or permanent changes, such as an additional machine or a considerable change of the product mix, the EPT and overtaking distribution have to be re-estimated to characterize the new workstation behavior. The throughput range in which accurate predictions are made will then gradually increase over time.

V. Conclusion

The proposed 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 and computational effort compared to a full-detail simulation model. The process time in the aggregate model, referred to as the EPT, is sampled from an EPT distribution that depends on the momentary WIP. The WIP-dependent EPT distribution includes semiconductor behavior such as integrated processing, and outage delays. The order in which lots are processed is modeled by means of a WIP-dependent overtaking distribution; lots entering the queue have a probability of overtaking other lots. Key to our approach is that the WIP-dependent EPT distribution and overtaking distribution are determined from arrival and departure events, measured at the operational workstation.

We have first validated the method using a simulation case of a workstation in which we vary the number of parallel machines, the number of integrated processes, the dispatching rule, and the variability of the process time and the interarrival time. We concluded that the mean and 95% quantile of the cycle time distribution can be accurately predicted (i.e., prediction errors are typically less than 10%) in a throughput region around the training level. For throughput levels higher than the training level, the predictions of the mean and 95% quantile of the cycle time are more accurate for multi-machine workstations than for a single machine workstation. For throughput levels lower than the training level, the 95% quantile prediction improves if the process time variability increases. Furthermore, we observed that for the priority dispatching rule, the throughput range for which accurate 95% quantile predictions were obtained is smaller than for FCFS and LCFS dispatching rules.

In a second experiment, we have investigated the effect of limiting the size of data set using a simulation model that may be viewed as a lithography workstation. In this experiment, we predicted the cycle time distribution using 10^2, 10^3, 10^4, and 10^5 measured arrivals and departures, respectively, to estimate the EPT distribution and overtaking distribution. We have introduced a curve fitting approach to overcome the difficulties with noise that arise because of the limited amount of data. We concluded that the mean and 95% quantile of the cycle time can be accurately predicted at the training level, independent of the number of measured arrival and departure events. The range of throughput ratios around the training level for which accurate predictions can be obtained increases for an increasing number of measured events. Additionally, the second experiment shows that the proposed method can accurately predict the cycle time when multiple product types are processed by the workstation.

For all simulation experiments, we have observed that the calculation time required to evaluate the aggregate simulation model is about 20 times shorter than the calculation time required to evaluate the detailed simulation model.

We have demonstrated the applicability of the proposed method in semiconductor practice by applying the method to a Crolles2 lithography workstation. The results show that the tail of the cycle time distribution is accurately predicted at the actual throughput level of operation. The results of the simulation test case suggest that accurate predictions can also be made for throughput levels other than the operational throughput.

The aggregate modeling method can be used for planning purposes to make a tradeoff between the throughput and the cycle time distribution of the workstation. Lithography is usually the main contributor to the cycle time of lots. We expect that the method can also be used for other semiconductor workstations, such as the metal or implant workstations. These workstations also have wafers of multiple lots in process at the same time.

The proposed aggregate model may be also be helpful in areas other than production planning. In their survey, Taylor and Robinson [24] stated that there is a need for higher level modeling techniques that abstract away from low-level model detail to justify the development of a detailed model. The aggregate model presented in this paper may be helpful in this respect. Furthermore, Fowler and Rose [25] stated that reducing problem solving cycles is a grand challenge in modeling and simulation of complex manufacturing systems. The aggregate model proposed in this paper can be developed much faster than a detailed simulation model.

In future research, we will show how the aggregate modeling concept can be used to build a model of an entire manufacturing network. The factory can be modeled as a network of aggregate servers of the type presented in this paper, where each aggregate server represents a workstation. Such a model could be used to predict the on-time delivery performance of the factory. In case the cycle time of each individual lot is important (which is referred to as pegging), the cycle times of lots processed in the aggregate model could be directly coupled to the due dates of these lots. In case customers are served in FCFS order, not caring which particular lot they receive (netting), the aggregate model could be used assuming FCFS, because the order in which lots are processed is not relevant. No overtaking is required in the model then; the model becomes the same as the single-server aggregate model presented in [20].

The EPT based aggregate model developed in this paper could be extended to distinguish between multiple product types; in the present model, all product types are aggregated into a single product type (see case II in Section III). To incorporate multiple product types in the aggregate model, the measured EPT and overtaking realizations could be assigned to product types, in addition to the assignment to WIP levels. This poses additional challenges for the data collection, because the measured EPT and overtaking realizations have to be spread over more groups.

In this paper, we focused on the prediction of the cycle time distribution. Another relevant research topic is to investigate how well the aggregate model captures correlations between consecutive cycle times.

APPENDIX

The algorithm used to calculate EPT-realizations and overtaking realizations is depicted in Fig. 12. The following variables are used: variable τ denotes the event time, variable ev the event type (either an arrival **a** or a departure **d**), and i the lot arrival number (so Lot i is the ith arriving lot). Furthermore, variable xs is a list that stores for each lot in the system its arrival number, i, and the number of lots in the system just before its arrival aw. Variable s is used to store the EPT start time. Variable sw stores the number of lots in the system just after 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 (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 just before arrival (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 just after 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 just before 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 just before the arrival of Lot i (aw) are written. If there are still lots in the system after the departure (len(xs) > 0), a new EPT start time is stored in s, as well as the corresponding number of lots currently in the system (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 just before 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, aw) 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
```

Fig. 12. EPT Algorithm (top) and function detOvert (bottom).

ACKNOWLEDGMENT

The authors would like to thank B. Lemmen of Crolles2 for his support in obtaining the data.

REFERENCES

[1] L. Kleinrock, *Queueing Systems. Volume I: Theory*, 1st ed. New York: Wiley, 1975.

[2] J. G. Shanthikumar, S. Ding, and M. T. Zhang, "Queueing theory for semiconductor manufacturing systems: A survey and open problems," *IEEE Trans. Autom. Sci. Eng.*, vol. 4, no. 4, pp. 513–522, Oct. 2007.

[3] P. Backus, M. Janakiram, S. Mowzoon, G. C. Runger, and A. Bhargava, "Factory cycle-time prediction with a data-mining approach," *IEEE Trans. Semicond. Manuf.*, vol. 19, no. 2, pp. 252–258, May 2006.

[4] Y. Hung and C. Chang, "Using a empirical queueing approach to predict future flow times," *Comput. Ind. Eng.*, vol. 37, no. 4, pp. 809–821, 1999.

[5] C. F. Chien, C. W. Hsiao, C. Meng, K. T. Hong, and S. T. Wang, "Cycle time prediction and control based on production line status and manufacturing data mining," in *Proc. Int. Symp. Semicond. Manuf.*, 2005, pp. 327–330.

[6] A. Raddon and B. Grigsby, "Throughput time forecasting model," in *Proc. IEEE/SEMI Adv. Semicond. Manuf. Conf.*, Sep. 1997, pp. 430–433.

[7] J. M. Bekki, J. W. Fowler, G. T. Mackulak, and B. Nelson, "Indirect cycle-time quantile estimation using the Cornish-Fisher expansion," *IIE Trans.*, vol. 42, no. 1, pp. 31–44, 2010.

[8] J. M. Bekki, G. T. Mackulak, and J. W. Fowler, "Indirect cycle-time quantile estimation for non-FIFO dispatching policies," in *Proc. Winter Simulation Conf.*, Dec. 2006, pp. 1829–1835.
[9] A. I. Sivakumar and C. S. Chong, "A simulation based analysis of cycle time distribution, and throughput in semiconductor backend manufacturing," *Comput. Ind.*, vol. 45, no. 1, pp. 59–78, May 2001.
[10] W. Dangelmaier, D. Huber, C. Laroque, and M. Aufenanger. (2007). To automatic model abstraction: A technical review. *Proc. 21st Eur. Conf. Modeling Simulation* [CD-ROM].
[11] F. Yang, B. E. Ankenman, and B. L. Nelson, "Estimating cycle time percentile curves for manufacturing systems via simulation," *INFORMS J. Comput.*, vol. 20, no. 4, pp. 628–643, 2008.
[12] E. J. Chen, "Metamodels for estimating quantiles of systems with one controllable parameter," *SIMULATION*, vol. 85, no. 5, pp. 307–317, 2009.
[13] R. J. Brooks and A. M. Tobias, "Simplification in the simulation of manufacturing systems," *Int. J. Prod. Res.*, vol. 38, no. 5, pp. 1009–1027, 2000.
[14] R. T. Johnson, J. W. Fowler, and G. T. Mackulak, "A discrete event simulation model simplification technique," in *Proc. Winter Simulation Conf.*, 2005, pp. 2172–2176.
[15] O. Rose, "Why do simple wafer fab models fail in certain scenarios?" in *Proc. Winter Simulation Conf.*, 2000, pp. 1481–1490.
[16] O. Rose, "Improved simple simulation models for semiconductor wafer factories," in *Proc. Winter Simulation Conf.*, 2007, pp. 1708–1712.
[17] W. J. Hopp and M. L. Spearman, *Factory Physics: Foundations of Manufacturing Management*, 3rd ed. New York: IRWIN/McGraw-Hill, 2008.
[18] J. H. Jacobs, L. F. P. Etman, E. J. J. van Campen, and J. E. Rooda, "Characterization of operational time variability using effective process times," *IEEE Trans. Semicond. Manuf.*, vol. 16, no. 3, pp. 511–520, Aug. 2003.
[19] K. Wu and K. Hui, "The determination and indetermination of service times in manufacturing systems," *IEEE Trans. Semicond. Manuf.*, vol. 21, no. 1, pp. 72–82, Feb. 2008.
[20] A. A. A. Kock, L. F. P. Etman, J. E. Rooda, I. J. B. F. Adan, M. V. Vuuren, and A. Wierman, "Aggregate modeling of multi-processing workstations," Eurandom, Eindhoven, The Netherlands, Eurandom Rep. 2008-032, Aug. 2008 [Online]. Available: http://www.eurandom.nl/reports
[21] C. P. L. Veeger, L. F. P. Etman, J. van Herk, and J. E. Rooda, "Generating cycle time-throughput curves using effective process time based aggregate modeling," in *Proc. ASMC*, May 2008, pp. 127–133.
[22] I. Adan, M. V. Eenige, and J. Resing, "Fitting discrete distribution on the first two moments," *Probab. Eng. Inform. Sci.*, vol. 9, no. 4, pp. 623–632, 1995.
[23] A. Hofkamp and J. Rooda, *χ 1.0 Reference Manual*. Eindhoven, The Netherlands: Systems Engineering Group, Eindhoven University of Technology, 2007 [Online]. Available: http://se.wtb.tue.nl/sewiki/chi
[24] S. J. E. Taylor and S. Robinson, "So where to next? A survey of the future for discrete-event simulation," *J. Simulation*, vol. 1, pp. 1–6, Dec. 2006.
[25] J. W. Fowler and O. Rose, "Grand challenges in modeling and simulation of complex manufacturing systems," *SIMULATION*, vol. 80, no. 9, pp. 469–476, 2004.

C. P. L. Veeger received the M.S. and Ph.D. degrees from the Systems Engineering Group, Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands, in 2006 and 2010, respectively.

He is currently a consultant with OM Partners, Wommelgem, Belgium. His research work was on the development of the effective process time method in semiconductor manufacturing.

L. F. P. Etman is currently an Associate Professor with the Systems Engineering Group, Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. His current research interests include simulation-based optimization, multi-disciplinary design optimization, and the effective process time method for performance analysis of manufacturing systems.

E. Lefeber received the M.S. degree in applied mathematics and the Ph.D. degree in tracking control of nonlinear mechanical systems, both from the University of Twente, Enschede, The Netherlands, in 1996 and 2000, respectively.

Since 2000, he has been an Assistant Professor with the Systems Engineering Group, Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. His current research interests include modeling and control of manufacturing systems.

I. J. B. F. Adan received the M.S. and Ph.D. degrees in mathematics from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 1987 and 1991, respectively.

Since 2000, he has been an Associate Professor with the Stochastic Operations Research Group, Department of Mathematics and Computing Science, Eindhoven University of Technology. Since 2009, he also has been a part-time Full Professor with the Operations Research and Management Group, Department of Quantitative Economics, University of Amsterdam, Amsterdam, The Netherlands. His current research interests include the analysis of multi-dimensional Markov processes and queueing models, and the performance evaluation of communication, production, and warehousing systems.

J. van Herk is currently a Quality Assurance and Safe Launch Engineer with Business Line Automotive Safety and Comfort, NXP Semiconductors, Nijmegen, The Netherlands. His current research interests include the optimization of the quality of products, manufacturing effectiveness, and advanced equipment and process control.

J. E. Rooda is currently a Professor with the Systems Engineering Group, Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. His current research interests include design and analysis of manufacturing systems, manufacturing control, and supervisory machine control.