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

Predicting the cycle time distribution as a function of throughput is helpful in making a trade-off between workstation productivity and meeting due dates. To predict cycle time distributions, detailed models are almost exclusively used, which require considerable development and maintenance effort. Instead, we propose a so-called aggregate model to predict cycle time distributions, which is a lumped-parameter representation of the queueing system. The lumped parameters of the model are determined directly from arrival and departure events measured at the workstation. The paper demonstrates that the aggregate model can accurately predict the cycle time distribution of workstations in semiconductor manufacturing, in particular the tail of the distribution.

Index Terms

cycle time distribution, discrete-event simulation, queueing, manufacturing systems, performance evaluation, factory dynamics

I. Introduction

In production planning for semiconductor workstations, there is a trade-off between productivity and meeting due dates. With a workstation, we mean 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. A high workstation productivity is desirable because of the capital intensive equipment used. On the other hand, a high workstation productivity causes high cycle times, with cycle time defined as the sum of process time and waiting time at the workstation. As a consequence, a smaller percentage of lots will meet their due date.

To make a trade-off between productivity and due dates, 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. 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 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], [4], [5], [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, McNeill *et al.* [7] and Bekki *et al.* [8] estimated a set of quantiles from a detailed simulation model 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. 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

* Submitted for publication

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, The Netherlands (e-mail: {c.p.l.veeger,l.f.p.etman,a.a.j.lefeber,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, The Netherlands (e-mail: i.j.b.f.adan@tue.nl), and with the Operations Research and Management Group, Department of Quantitative Economics, University of Amsterdam

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

except the bottleneck station. He concluded that the proposed model inaccurately estimates cycle time distributions for certain scenarios. To improve the estimations, Rose [16] replaced the delay distributions by a FCFS (First-Come-First-Served) 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 workstation in full detail first. The proposed model is a single-server aggregate queueing model. The lumped parameters of the model are determined from arrival and departure data measured at the workstation in operation. We demonstrate that the aggregate model can accurately predict cycle time distributions of workstations in semiconductor manufacturing.

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. With 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. Since 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 integrated processing workstations, the EPT distribution parameters are typically WIP-dependent, because multiple lots may be in process at the same time. 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 First-Come-First-Served (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 be effectively used to predict cycle time distributions for semiconductor workstations. We first validate the method using a simulation test 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. In the simulation case, ample arrival and departure events are available. However, in semiconductor practice, typically a limited number of measured events are available. In a second simulation case representing a lithography workstation, we show how still accurate predictions can be made when a limited amount of data is available. Finally, a test case based on data from the Crolles2 factory demonstrates the applicability of the method in semiconductor manufacturing.

The outline of the 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. Figure 1a illustrates an integrated processing workstation, which consists of m identical parallel machines, each of which have l sequential integrated processes. Figure 1b visualizes the proposed aggregate model. In this section we introduce the aggregate model concept and explain how we determine model parameters.

A. The Aggregate Model

We propose the following aggregate model (Figure 1b). 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 i^{th} *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 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 $F_{\text{K}}(k; w)$, which defines the probability $P(K \leq k; w)$ that k or less lots are overtaken. Probability distribution $F_{\text{K}}(k; w)$ 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

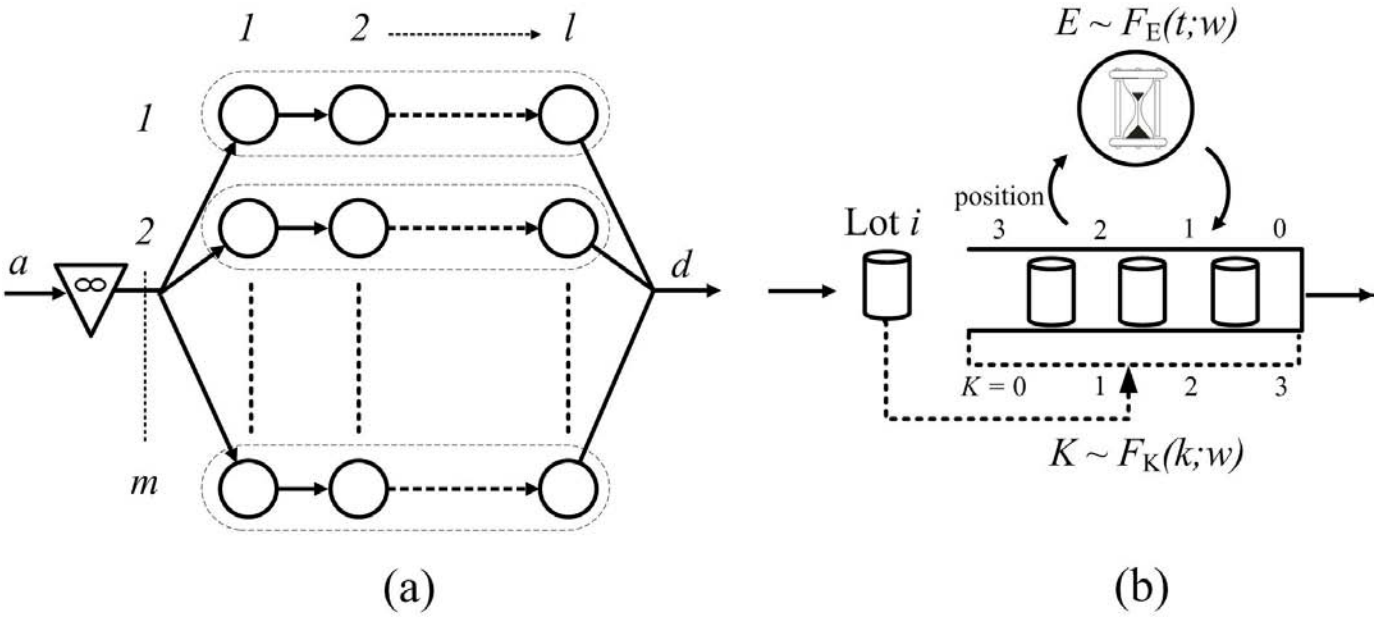


Fig. 1. An example of a workstation (a), and the proposed aggregate model (b).

the head of the queue. For example, in Figure 1b, $w = 3$ 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: i) a lot arrives while no lots are present in the queue, or ii) a lot departs while leaving one or more lots behind. When the timer starts, a time period E is sampled from probability distribution $F_{\mathrm{E}}(t; w)$, which defines the probability $P(E \leq t; w)$ that E is less than or equal to t. The probability distribution $F_{\mathrm{E}}(t; w)$ depends on number of lots w in the system just after the timer start. So in case of a lot arrival (case i)), w includes the arrived lot. In case of a lot departure (case ii)), w does not include the departed lot. Time period E is referred to as an Effective Process Time (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 $F_{\mathrm{E}}(t; w)$ per WIP-level w and an overtaking distribution $F_{\mathrm{K}}(k; w)$ per WIP-level w. We assume that all sampled EPT realizations, and overtaking realizations are independent.

B. Example

Figure 2a shows four lots processed by the aggregate model in FCFS order. The first row of Figure 2a 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 Figure 2a 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 Figure 2a 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). Here, 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. Figure 2a shows that for each arrival the sampled number of overtaken lots equals zero, which implies that no overtaking occurs; the order of arrival is equal to the order of departure.

Figure 2b shows four lots with overtaking. The lot arrival times, and the sampled EPTs are the same as in Figure 2a, 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 (Lot 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 i^{th} *arriving* lot) departing

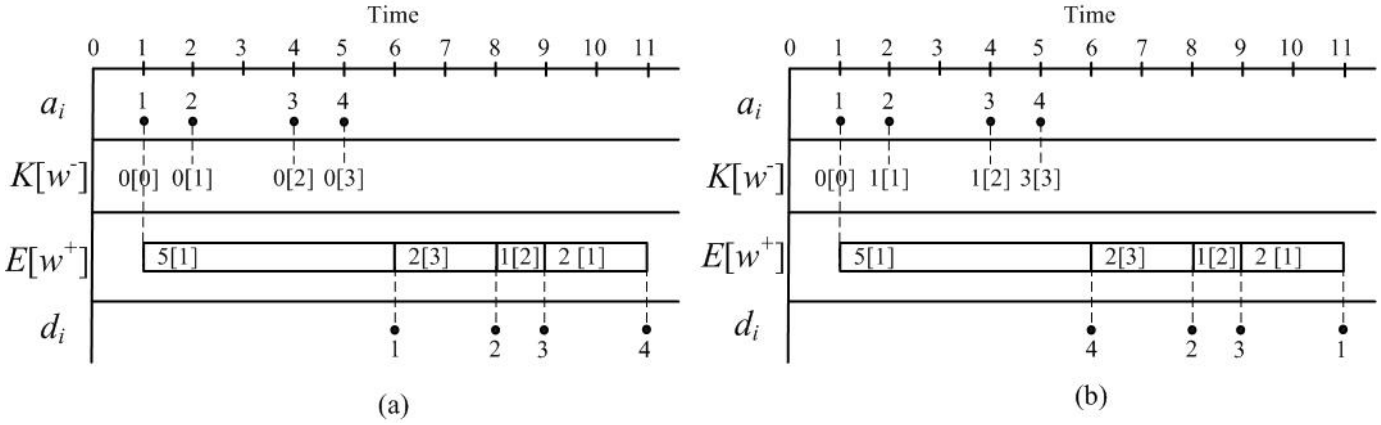


Fig. 2. Lot-time diagrams of four lots processed by the aggregate model including the EPTs sampled by the timer and the sampled number of overtaken lots; (a) without overtaking, (b) with overtaking.

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 EPT realizations, the number of lots overtaken by each lot, as well as the corresponding WIP-levels are determined 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 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 duration of the EPT by subtracting the EPT start time from the departure time (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. 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 output.

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 a distribution is obtained, which is used in the aggregate model for the EPT distribution $F_{\mathrm{E}}(t; w)$ of the corresponding WIP-level.

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

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. The second case is used to investigate the predictions for a workstation representing a lithography workstation for which a limited amount of measured arrival and departure events is available.

A. Case I

1) Description: Case I is depicted in Figure 1a. The workstation consists of m identical parallel machines. Each machine consists of l sequential integrated processes that each have a gamma-distributed process time with mean t_0 and coefficient of variation c_0. Lots arrive in 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).

We experiment with different values of m, l, c_0, and c_a. For the dispatching rule d, we consider First-Come-First-Served (FCFS), non-preemptive Last-Come-First-Served (LCFS), and Priority (Pr) dispatching. For FCFS and LCFS dispatching, we assume that all lots require the same process time $t_0 = 1.0$, and c_0 in the various processes. For Pr dispatching, we use two

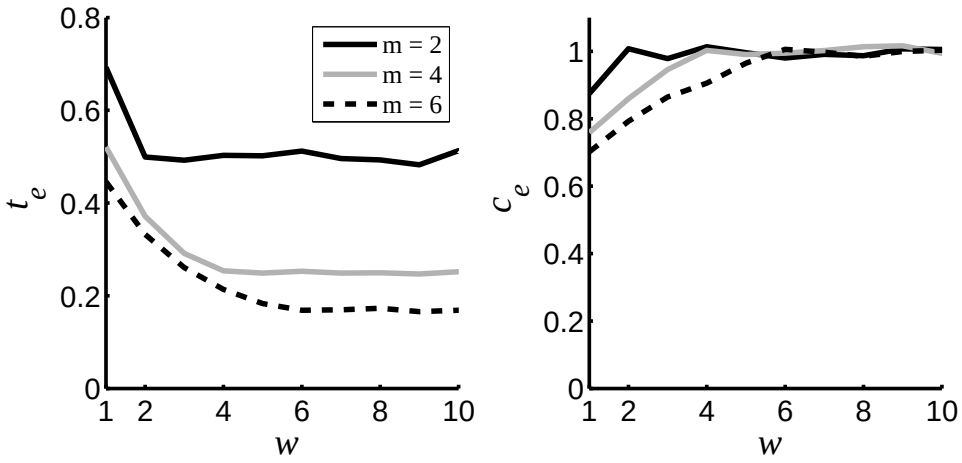


Fig. 3. Mean EPT t_e and CV c_e as a function of WIP level w for Case I for different values of m, and constant $l = 1$, $c_0 = c_a = 1.0$, and $d =$ FCFS

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) Calculating model parameters: To obtain the WIP-dependent EPT distribution $F_{\mathrm{E}}(t; w)$ and overtaking distribution $F_{\mathrm{K}}(k; w)$ for a workstation configuration, arrivals and departures of 10^5 lots were obtained 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. To represent the EPT distribution for each WIP-level, we use a gamma distribution with mean t_e and coefficient of variation c_e. Distribution parameters t_e and c_e are directly obtained from the measured data. Recall that maximum WIP-level $w_{\max}$ groups all EPTs that started with WIP-level $w \geq w_{\max}$. On the one hand, we want to choose $w_{\max}$ as high as possible to include system behavior at high WIP-levels. On the other hand, the higher $w_{\max}$, the more difficult it becomes to accurately estimate the EPT distribution parameters for high WIP-levels, because we typically obtain little EPT measurements at high WIP levels. We therefore choose $w_{\max}$ as high as possible, under the condition that the 95% confidence interval of $t_{e,w_{\max}}$ is less than $\pm$ 2.5%.

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.

To illustrate the proposed method, we now present the measured EPT distribution parameters and the measured overtaking probabilities for a selection of workstation configurations. Figure 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, 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 $w = m$. For this configuration, c_e increases for increasing w, until $w \approx m$ where c_e reaches 1.0. For low $w < m$, the interdeparture time between lots depends on the exponential arrival process and the exponential service process, whereas for $w \geq m$ the interdeparture time only depends on the exponential service process.

Next we show that the overtaking distribution $F_{\mathrm{K}}(k; w)$ depend on the dispatching rule. Figure 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 $w > 1$. This is because 50% of the arriving lots is of type B (with long process times), which do not overtake. The type A may overtake one or more type B lots in the queue, 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 $k = w - 1$.

Figure 5 shows that the overtaking probabilities depend on c_0. In Figure 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 Figure 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.

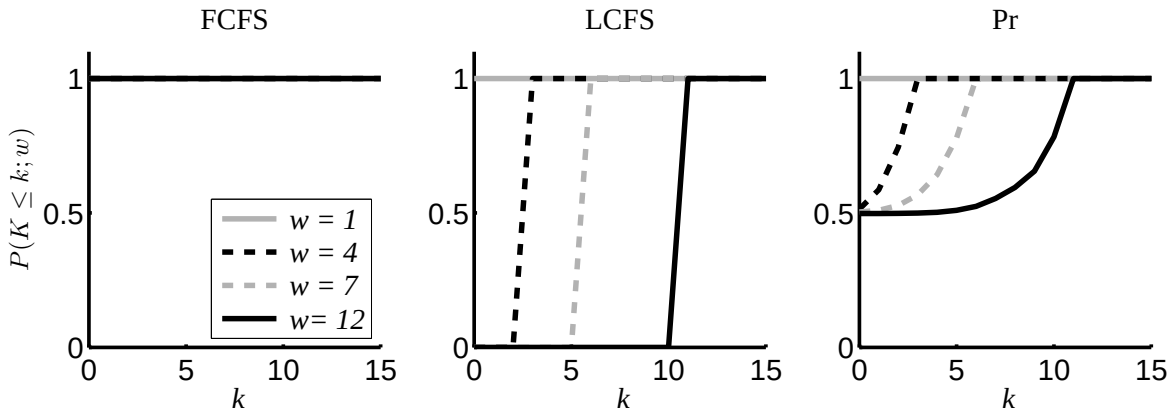


Fig. 4. Case I: cumulative probability for a newly arrived lot to overtake K lots already in the system for various WIP-levels w and for different dispatching rules, with $m = l = 1$, and $c_0 = c_a = 1.0$.

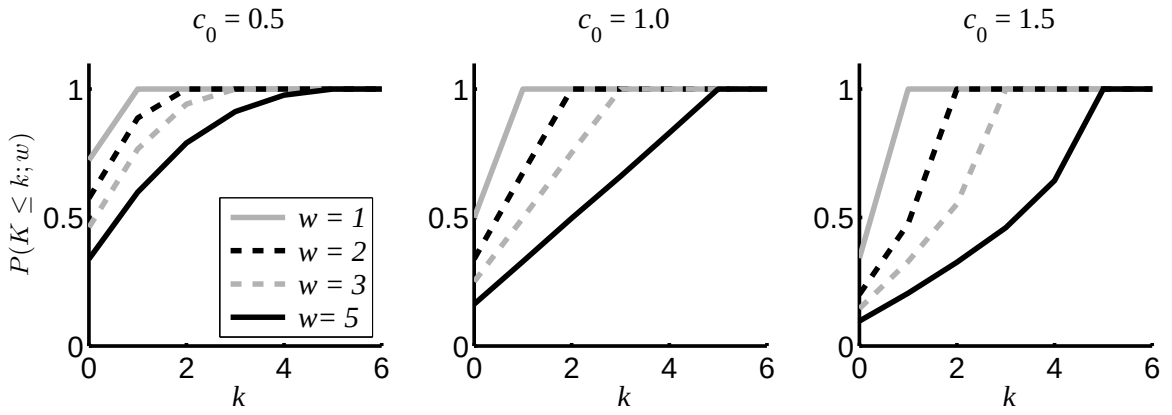


Fig. 5. Case I: cumulative probability for a newly arrived lot to overtake K lots already in the system for various WIP-levels w and for different values of arrival coefficient of variability c_0, with $m = 6$, $l = 1$, $c_a = 1.0$, and $d = \text{FCFS}$.

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. This is 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 $\delta/\delta_{\max}$ of 0.6, 0.8, and 0.9. For each throughput ratio, 15 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 Figure 1b to predict cycle time distributions. The aggregate model is trained at $\delta/\delta_{\max} = 0.8$ using 10^5 arrivals and departures generated using 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 15 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.

Figure 6 depicts cycle time distributions of the workstation (the black lines), and cycle time distributions predicted by the aggregate model (the dashed grey lines) for workstation configurations with $c_0 = \{0.5, 1.0, 1.5\}$, with $m = 6$, $l = 3$, $d = \text{FCFS}$, 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

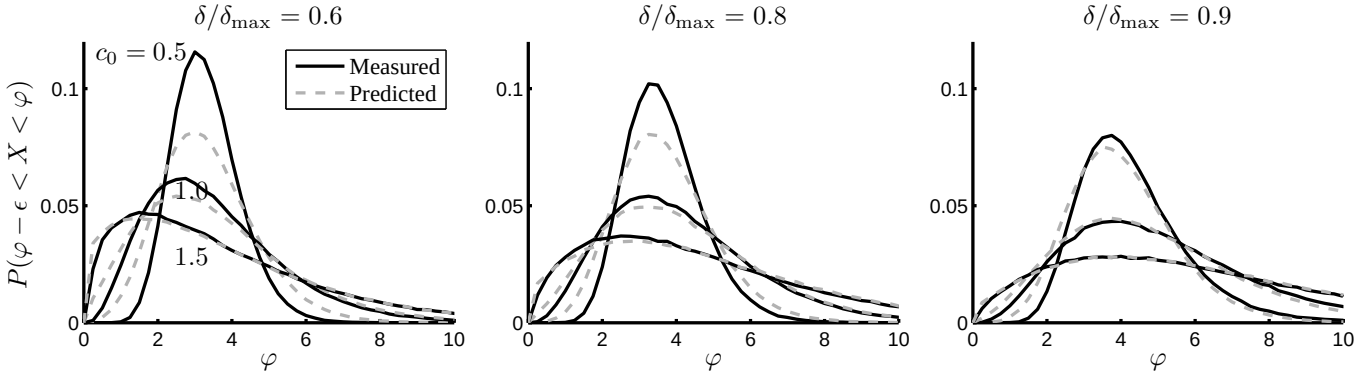


Fig. 6. Flow time distribution of the considered workstation, and predicted by the proposed aggregate model for $c_0 = \{0.5, 1.0, 1.5\}$, with $m = 6$, $d = \text{FCFS}$, $l = 3$, and $c_\text{a} = 1.0$.

training level. The different lines in the plots correspond to different values of c_0: the top solid black and dashed grey lines in each plot corresponds to $c_0 = 0.5$, the middle lines correspond to $c_0 = 1.0$, and the bottom lines 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.

Figure 6 shows that for all considered values of c_0 and $\delta/\delta_{\max}$, the tail of the cycle time distribution is predicted very accurately. For $c_0 = 1.0$ and 1.5, even the whole distribution is accurately predicted. For $c_0 = 0.5$ and $\delta/\delta_{\max} = 0.6$ and 0.8 predictions are less accurate for low cycle times. The measured cycle time distribution shows less variability than the predicted cycle time distribution. This may be due to the fact that the EPT and the number of overtaken lots in the aggregate model are sampled independently for successive lots, possibly creating more variability than in reality.

We have also considered the cases $m = 1, 2$ and 4, using the same constant parameter values as in Figure 6. The accuracy of the predictions of the cycle time distribution is similar to the accuracy of the $m = 6$ case.

Figure 7 depicts cycle time distributions of the workstation, and cycle time distributions predicted by the aggregate model for workstation configurations with different values of l and d, with $m = 1$, and $c_0 = c_\text{a} = 1.0$. Figure 7a considers $d = \text{FCFS}$, Figure 7b depicts the results of $d = \text{LCFS}$, and Figure 7c depicts $d = \text{Pr}$. From left to right the plots consider different throughput ratios. The different lines in the plot represent different values of l, which are 2, 4, and 8. The top line represents $l = 2$, the middle line $l = 4$, and the bottom line $l = 8$.

Figure 7 shows that for all considered values of l, d, and $\delta/\delta_{\max}$, the tail of the cycle time distribution is predicted accurately. For $l = 2$, the whole distribution is predicted accurately. For increasing l, the predictions deteriorate for low cycle times. As in Figure 6, the measured cycle time distribution shows less variability than the predicted cycle time distribution. We expect this is also due to the fact that the EPT and the number of overtaken lots in the aggregate model are sampled independently for successive lots.

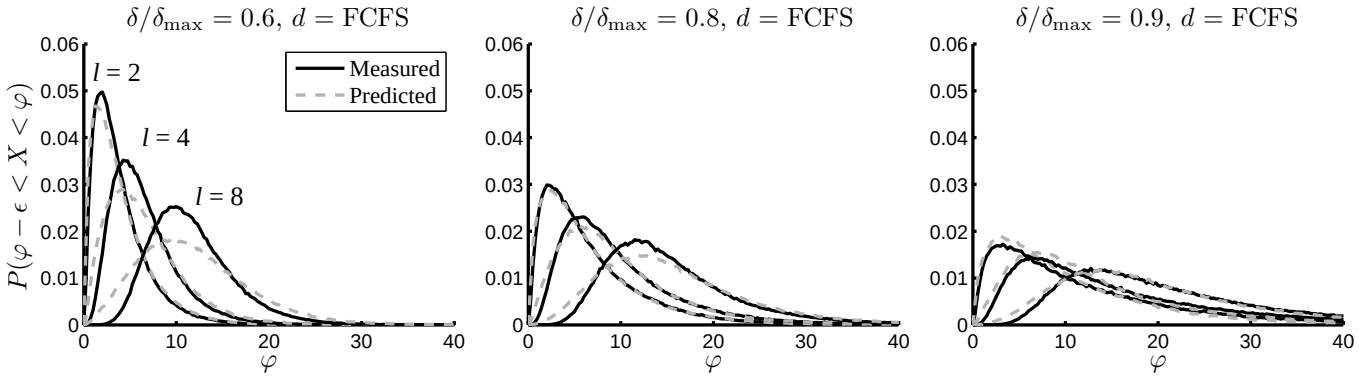
We have also experimented with different values of c_a (0.5 and 1.5). We observe that c_a has little influence on the accuracy of the cycle time predictions, since we also use c_a for the arrival process in the aggregate model.

B. Case II

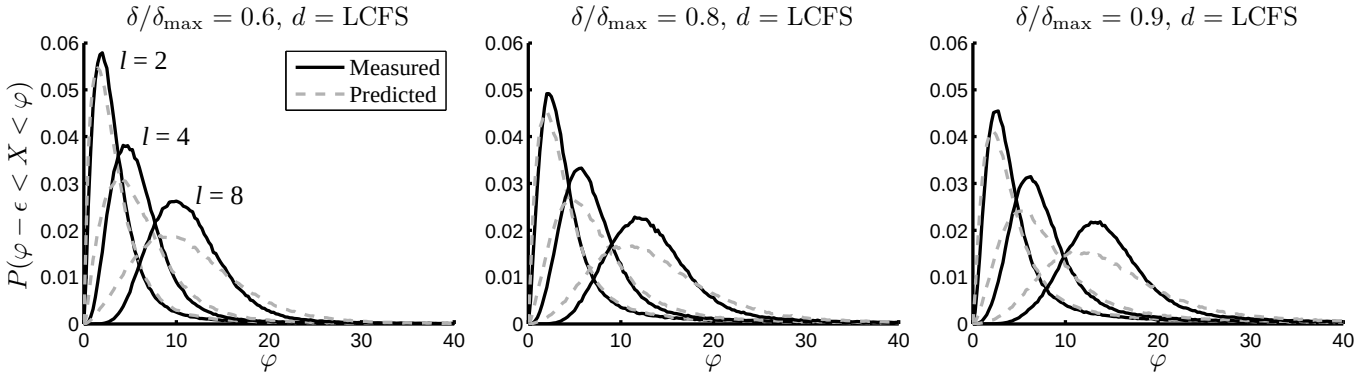
1) Description: Case II is depicted in Figure 8. The setup of Case II may be viewed as a group of track-scanner lithography tools. Lots arrive at the infinite buffer according to a Poisson process: 50% of the arriving lots is of type A, whereas the other 50% is of type B. Lots are processed in First-Come-First-Serve order taking into account machine recipe qualification. 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. 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 is 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.

2) Calculating model parameters: Arrivals and departures of 20000 lots were obtained at a throughput ratio of $\delta/\delta_{\max}$ of 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 again use gamma distributions to represent the EPT distributions for each WIP-level.

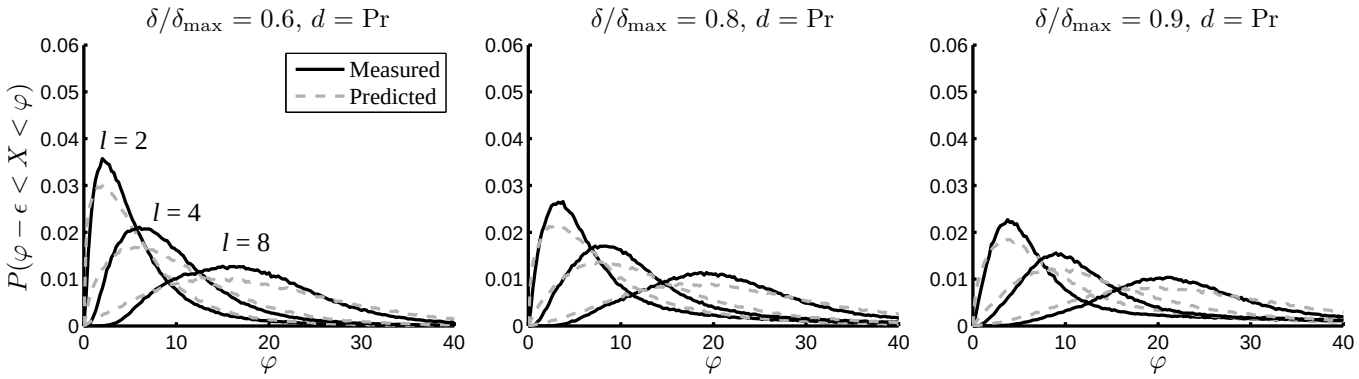
Other than for Case I, we now measure arrivals and departures of only 20000 lots, a number one may encounter in semiconductor manufacturing practice (see also Section IV). As a consequence, it is more difficult to accurately estimate



(a)



(b)



(c)

Fig. 7. Flow time distribution of the considered workstation, and predicted by the proposed aggregate model for $l = \{2, 4, 8\}$, with $m = 1$, and $c_a = c_0 = 1.0$: (a) considers FCFS, (b) LCFS, and (c) Pr dispatching.

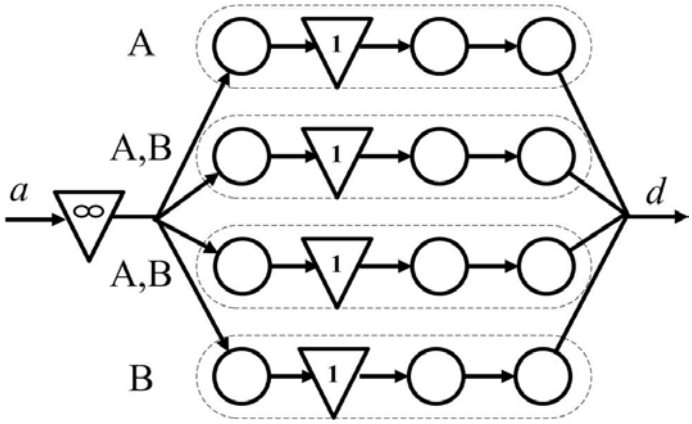


Fig. 8. Case II representing a lithography workstation

model parameters $t_{\mathrm{e}}(w)$, $c_{\mathrm{e}}(w)$, and $F_{\mathrm{K}}(k;w)$, and the cycle time predictions may deteriorate. In particular, we observe that an accurate estimation 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 $t_{\mathrm{e}}(w_{\max})$ estimation, 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}. Furthermore, for WIP-levels smaller than $w_{\max}$ we observe noise on $t_{\mathrm{e}(w)}$ and $c_{\mathrm{e}(w)}$ because little EPT realizations may have been collected for certain WIP-levels. To overcome this problem of noise we introduce a curve fitting approach, similar to [21].

The left plot in Figure 9 shows $t_{\mathrm{e}}(w)$ (the black line), obtained using only 20000 arrivals and departures. The middle plot of Figure 9 shows $c_{\mathrm{e}}(w)$ (the black line). We choose $w_{\max} = 15$, because for $w > 15$, $t_{\mathrm{e}}(w)$ does not decrease further. Note that Figure 9 shows noise on the values of $t_{\mathrm{e}}(w)$ and $c_{\mathrm{e}}(w)$ for $w < 15$. To deal with the noise, we approximate the measured $t_{\mathrm{e}}(w)$ values by $\hat{t}_{\mathrm{e}}(w)$, for which we use the following exponential function [21]:

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

Herein, θ represents the value of $\hat{t}_{\mathrm{e}}(w)$ at $w = \infty$. Variable η represents the value of $\hat{t}_{\mathrm{e}}(w)$ at $w = 1$. Variable λ represents the 'decay constant' of the exponential curve. We set η equal to the measured $t_{\mathrm{e}}(w)$ value for $w = 1$. We set θ such that $\hat{t}_{\mathrm{e}}(15)$ is equal to the measured t_{e} for $w = 15$. 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 the measured c_{e} values by $\hat{c}_{\mathrm{e}}(w)$, which is of the same exponential form as Equation (1). The obtained values of θ, η, and λ are 0.5899, 0.8265, and 1.0413 respectively.

The right plot of Figure 9 shows the cumulative overtaking probabilities $P(K \leq k;w)$ as a function of k for several values of w. For the overtaking distribution, we do not introduce a curve fit; we use the measured overtaking distribution directly in the aggregate model. For WIP-levels for which we have not measured the overtaking probabilities, we again assume that the overtaking probabilities are the same as for the highest measured WIP-level. In principle, a curve fit could be used to represent the overtaking probabilities. For example, [22] fit discrete distributions for which the stochastic variable exists in the range $[0, 1, .., \infty]$. However, in our case a sampled K value is always less than or equal to a finite value (w). For this type of distribution, little results are reported.

3) Cycle time predictions: The detailed simulation model of the Case II workstation is used to calculate the real cycle time distribution for throughput ratios $\delta/\delta_{\max}$ of 0.6, 0.8, and 0.9. 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 for Case I.

The aggregate model depicted in Figure 1b is used to predict cycle time distributions. 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 the same as 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 $\hat{t}_{\mathrm{e}}(w)$ and coefficient of variability $\hat{c}_{\mathrm{e}}(w)$. For the overtaking distributions in the aggregate model, we use the empirical overtaking distributions (as we did in Case I).

Figure 10 depicts the cycle time distributions obtained for the considered workstation and the aggregate model at throughput ratios 0.6, 0.8, and 0.9. The figure shows that again the tail of the cycle time distribution is accurately predicted, even though we have measured only 20000 arrivals and departures. For low cycle times, predictions are less accurate because of the effect explained in Section III-A.

IV. CROLLES2 CASE

We finally apply the proposed method to an operational workstation at the Crolles2 wafer fab. Crolles2 is a multi-product 300mm fab in which both high volume products and small series and prototype products are produced. Standard production lots,

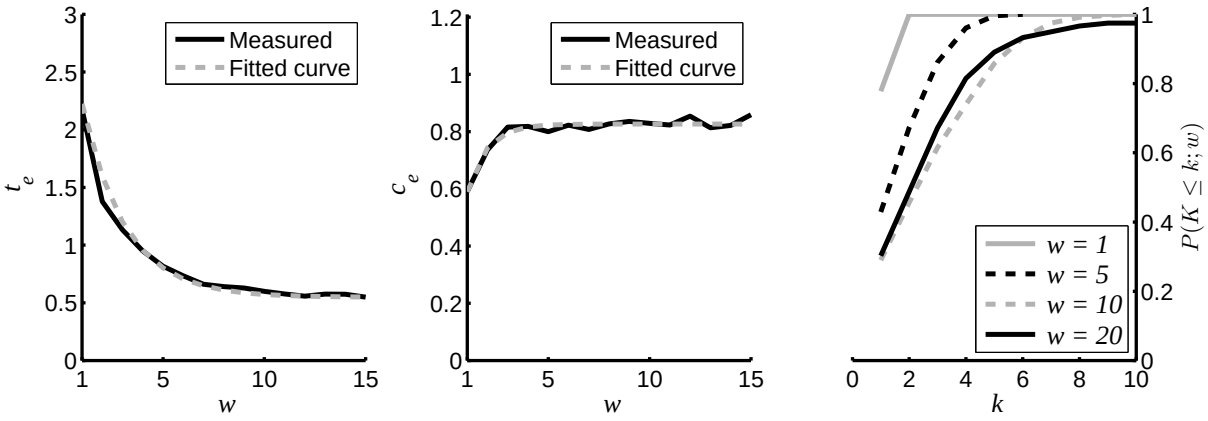


Fig. 9. Measured mean EPT t_e (left) and coefficient of variability c_e (middle) with fitted curves, and measured cumulative overtaking probabilities (right) of Case II using 20000 arrivals and departures

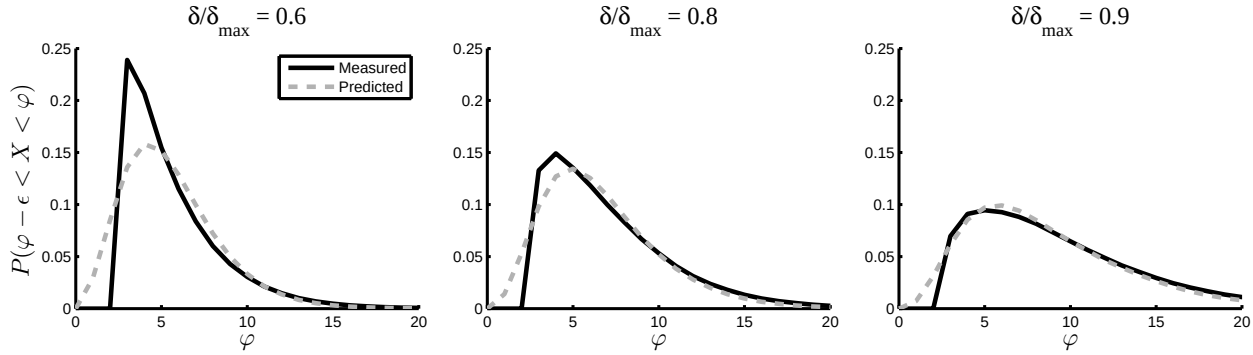


Fig. 10. Cycle time distribution of Case II, and predicted by the aggregate model using 20000 arrivals and departures for throughput ratio 0.6, 0.8, and 0.9

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 distributions. Finally, cycle time distributions are predicted using the aggregate model.

A. 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.

B. Calculating model 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 is filtered as described in [21]. After this filtering, the EPT algorithm in Appendix A is used to calculate EPT-realizations and lot overtaking realizations. We choose $w_{\max} = 100$; 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 Figure 11 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

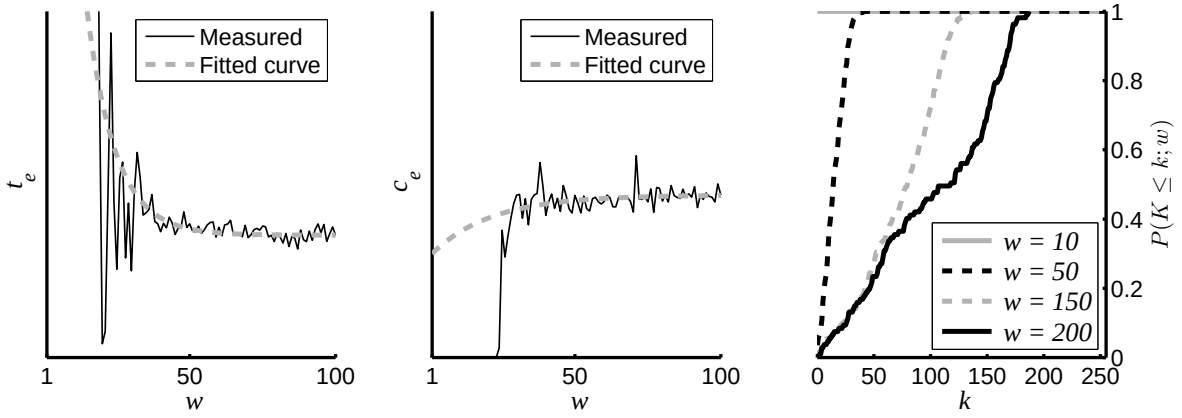


Fig. 11. Measured mean EPT t_e (left) and coefficient of variability c_e (middle) with fitted curves, and measured cumulative overtaking probabilities (right) of the Crolles2 lithography workstation

procedure described in Section III using exponential function (1). Note that we do not have EPT realizations for $w < 18$. For t_e and c_e at $w = 1$ we estimate values; t_e and c_e at $1 < w < 18$ then follow from the curve fit.

The left plot of Figure 11 clearly illustrates that the mean interdeparture 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 11 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 lower than 18 and 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 Figure 1b to estimate cycle time distributions of the lithography workstation, using gamma-distribution EPT distributions based on fitted values $\hat{t}_e(w)$ and $\hat{c}_e(w)$, and the empirical overtaking overtaking distribution as model parameters. We again perform 15 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.

Figure 12 depicts 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 training point δ^*. We use here the relative throughput instead of throughput ratio $\delta/\delta_{\max}$ because of 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.

Figure 12 shows that the cycle time distribution is accurately estimated at the training 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 training point. The simulation test cases described in Section III indicates that accurate predictions can be made for throughput levels other than the training point, in particular for the tail of the distribution. Therefore, we expect that accurate cycle time distributions can be obtained at throughput levels other than the training point.

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 Effective Process Time (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 buffer have a probability to overtake 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.

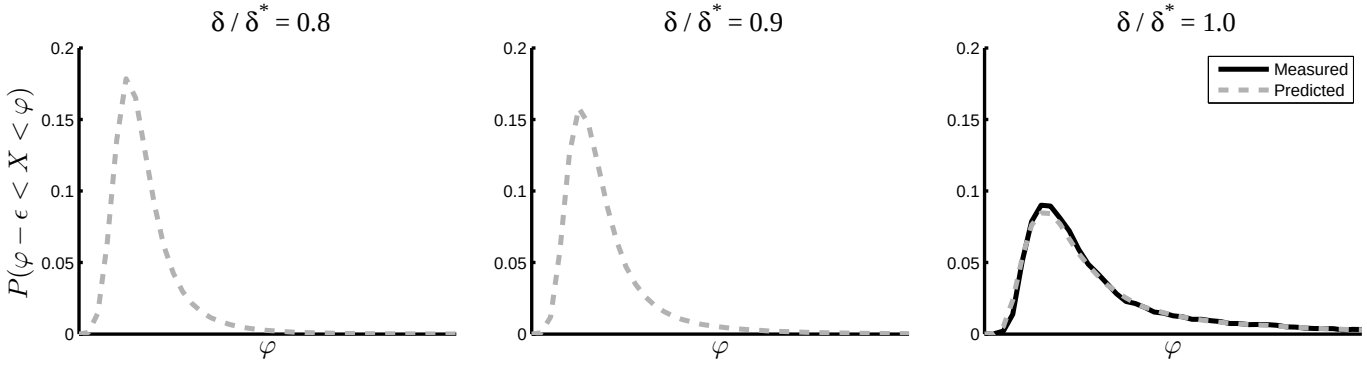


Fig. 12. Cycle time distribution of the litho workstation, and predicted by the proposed method for relative throughput levels of 0.8, 0.9, and 1.0

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 conclude that cycle time distributions can be accurately predicted by the aggregate model, in particular the tail of the distribution.

In a second experiment, we have investigated the effect of a limited data set using a simulation model that may be viewed as a lithography workstation. In this experiment, we measured only 20000 arrivals and departures to estimate the EPT distribution and overtaking distribution, which is an amount that may typically be encountered in semiconductor practice. 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 an accurate cycle distribution prediction when comparing the simulated cycle time distributions with the measured cycle time distribution at the throughput level of operation. The results of the simulation test case suggest that also accurate predictions can be made for throughput levels other than the operational throughput.

In this way, for planning purposes, the maximum throughput can be estimated for which 95% of the lots are completed within a user-defined time span using the 95% quantile. Lithography is usually the main contributor to the cycle time of lots. We expect that the proposed 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 [23] state 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. Furthermore, Fowler and Rose [24] state that reducing problem solving cycles is a grand challenge in modeling and simulation of complex manufacturing systems. Abstract models such as the aggregate model proposed in the current paper may be helpful in these respects.

In future research we aim to investigate whether we can use our aggregate model to aggregate entire manufacturing networks.

ACKNOWLEDGMENTS

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

APPENDIX

The algorithm used to calculate EPT-realizations and overtaking realizations is depicted in Figure 13. The following variables are used: variable τ denotes the event time, variable ev the event type (arrival $\mathbf{a}$ or departure $\mathbf{d}$), and i the lot arrival number (so lot i is the i^{th} 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 ($\text{len}(xs) = 0$). 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 just before 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 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 ($\text{len}(xs) > 0$), 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 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. 13. EPT Algorithm (left) and function detOvert (right).

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, 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, 2006.

[4] Y. Hung and C. Chang, "Using a empirical queueing approach to predict future flow times," *Computers and industrial Engineering*, vol. 37, 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. Semiconduct. Manufact.*, 2005, pp. 327–330.

[6] A. Raddon and B. Grigsby, "Throughput time forecasting model," in *proc. IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 1997, pp. 430–433.

[7] J. E. McNeill, G. T. Mackulak, and J. W. Fowler, "Indirect estimation of cycle time quantiles from discrete event simulation models using the cornish-fisher expansion," in *Proceedings of the 2003 winter simulation conference*, December 2003, pp. 1377–1382.

[8] J. M. Bekki, G. T. Mackulak, and J. W. Fowler, "Indirect cycle-time quantile etimation for non-fifo dispatching policies," in *Proceedings of the 2006 winter simulation conference*, December 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," *Computers in Industry*, vol. 45, pp. 59–78, 2001.

[10] W. Dangelmaier, D. Huber, C. Laroque, and M. Aufenanger, "To automatic model abstraction: a technical review," in *proc. 21st European Conference on Modeling and Simulation*, 2007.

[11] F. Yang, B. E. Ankenman, and B. L. Nelson, "Estimating cycle time percentile curves for manufacturing systems via simulation," *INFORMS Journal on Computing*, vol. 20, no. 4, pp. 628–643, Fall 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," *International Journal of Production Research*, 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. 2005 Winter Simulation Conference*, 2005, pp. 2172–2176.

[15] O. Rose, "Why do simple wafer fab models fail in certain scenarios?" in *Proceedings of the 2000 Winter Simulation Conference*, 2000, pp. 1481–1490.
[16] O. Rose, "Improved simple simulation models for semiconductor wafer factories," in *Proceedings of the 2007 Winter Simulation Conference*, 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, 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," http://www.eurandom.tue.nl/reports, Eurandom report nr. 2008-032, August 2008.
[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 *Proceedings of the 2008 Advanced Semiconductor Manufacturing Conference (ASMC)*, May 2008, pp. 127–133, revised and extended version submitted to IEEE Trans. Semicond. Manuf.
[22] I. Adan, M. V. Eenige, and J. Resing, "Fitting discrete distribution on the first two moments," *Probability in the Engineering and Information Sciences*, vol. 9, pp. 623–632, 1995.
[23] S. J. E. Taylor and S. Robinson, "So where to next? a survey of the future for discrete-event simulation," *Journal of Simulation*, vol. 0, pp. 1–6, 2006.
[24] J. W. Fowler and O. Rose, "Grand challenges in modeling and simulation of complex manufacturing systems," *SIMULATION*, vol. 80, pp. 469–476, 2004.