Generating cycle time-throughput curves using effective process time based aggregate modeling

Citation for published version (APA):

Veeger, C. P. L., Etman, L. F. P., Herk, van, J., & Rooda, J. E. (2010). Generating cycle time-throughput curves using effective process time based aggregate modeling. *IEEE Transactions on Semiconductor Manufacturing*, 23(4), 517-526. <https://doi.org/10.1109/TSM.2010.2065490>

DOI:

[10.1109/TSM.2010.2065490](https://doi.org/10.1109/TSM.2010.2065490)

Document status and date:

Published: 01/01/2010

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Generating Cycle Time-Throughput Curves Using Effective Process Time Based Aggregate Modeling

C. P. L. Veeger, L. F. P. Etman, J. van Herk, and J. E. Rooda, *Member, IEEE*

Abstract—In semiconductor manufacturing, cycle time-throughput (CT-TH) curves are often used for planning purposes. To generate CT-TH curves, detailed simulation models or analytical queueing approximations may be used. Detailed models require much development time and computational effort. On the contrary, analytical models, such as the popular closed-form $G/G/m$ queueing expression, may not be sufficiently accurate, in particular, for integrated processing equipment that have wafers of more than one lot in process. Recently, an aggregate simulation model representation of workstations with integrated processing equipment has been proposed. This aggregate model is a $G/G/m$ type of system with a workload-dependent process time distribution, which is obtained from lot arrival and departure events. This paper presents a first proof of concept of the method in semiconductor practice. We develop the required extensions to generate CT-TH curves for workstations in a semiconductor manufacturing environment where usually only a limited amount of arrival and departure data is available. We present a simulation and an industry case to illustrate the proposed method.

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

I. INTRODUCTION

THE CYCLE time-throughput (CT-TH) curve of a workstation represents the relation between the mean cycle time of lots processed at the workstation and the throughput of the workstation. The cycle time is defined here as the queue time plus the process time of a lot at a workstation. With throughput we mean the number of lots processed per time unit at the workstation. CT-TH curves are used for production planning of semiconductor bottleneck workstations (e.g., in the litho area) to make a tradeoff between the workstation's utilization and the cycle time performance.

To calculate the CT-TH curve, $G/G/m$ queueing approximation (1) (see [1]–[3]) is often used in practice as follows:

$$\varphi = \left(\frac{c_a^2 + c_e^2}{2} \right) \left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)} \right) t_e + t_e \quad (1)$$

Manuscript received July 8, 2008; revised August 3, 2009; accepted May 19, 2010. Date of publication August 16, 2010; date of current version November 3, 2010.

C. P. L. Veeger, L. F. P. Etman, 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; j.e.rooda@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.2010.2065490

where mean cycle time φ is the sum of queue time (first term) and process time (second term). The mean queue time of a lot in a workstation with m parallel machines increases linearly with the squared coefficients of variation of interarrival time c_a^2 and process time c_e^2 , nonlinearly with utilization u , and linearly with mean process time t_e . Utilization u equals $u = \frac{t_e}{mt_a}$ with t_a the mean interarrival time. $G/G/m$ queueing approximation (1) assumes that when a machine is processing, its production rate does not depend on the workload.

Many machines in semiconductor manufacturing are integrated processing tools [4]. For these machines, the production rate increases when more lots are available for processing. Integrated processing tools carry out a series of process steps at various chambers inside the tool. Typically multiple loadports are present. When the workload increases, the machine can work on wafers of multiple lots at the same time, increasing its production rate. Chamber qualification for the various tools also contributes to this effect; the more lots are available for processing, the more likely it is that a lot is present for which the machine is qualified. Hence, the $G/G/m$ approximation (1) may be an inaccurate model representation for integrated processing tools.

Recently, a new aggregate queueing model has been proposed by Kock *et al.* [5], [6]. This aggregate model is a $G/G/m$ -like approximation but with a workload-dependent process time distribution. This process time distribution is denoted as the effective process time (EPT) distribution [3], [7].

The EPT can be seen as “the time seen by a lot at a workstation from a logistical point of view” [3]. The EPT includes not only the raw processing time but also all time losses due to down, setup, and any other source of variability. Hopp and Spearman [3] calculate the mean and variance of the EPT from these preemptive and non-preemptive outages, and use them in queueing approximation (1). Jacobs *et al.* [7] calculate t_e and c_e^2 directly from lot arrival and departure events.

Kock *et al.* [5], [6] also calculate the EPT distribution directly from arrival and departure events, but additionally label each EPT realization with the momentary workload. They use their aggregate model with the measured workload-dependent EPT distribution as input to generate a CT-TH curve. They tested their aggregate model for four flow line configurations in a simulation environment. Accurate CT-TH curves were obtained for these academic examples.

In this paper, we present a proof of concept of the method in semiconductor practice. We demonstrate that the method is able to generate CT-TH curves for workstations in an operating semiconductor environment. We develop the required extensions to handle the typically limited amount of arrival and departure data that can be obtained from the factory floor. (The simulation study in [5] and [6] uses millions of arrivals and departures, which is unrealistic in semiconductor practice.)

We start from track-in and track-out data obtained from the manufacturing execution system (MES). We then process this data such that a consistent set of lot arrivals and lot departures is extracted from the MES data. Following [5] and [6], we calculate an EPT-realization for each departing lot, but additionally register whether the EPT started upon arrival of the lot or upon the departure of the previous lot. To account for the limited number of lot arrivals and departures in the data collection period, we introduce a curve fitting approach. In this way, we obtain empirical relations for t_e and c_e^2 as a function of the workload.

We first illustrate the approach for a simulation example of a workstation with integrated processing equipment and investigate the accuracy of the CT-TH curve generated by the aggregate model. Next, we demonstrate the approach for four operational workstations in the Crolles2 semiconductor manufacturing facility (located in Crolles, France): lithography (track-scanners), metal, implant, and critical dimension measurement. We calculate CT-TH curves for all four workstations and compare them with the mean cycle time observed at the workstation during the data collection period.

The outline of this paper is as follows. In Section II, a brief overview of methods to calculate CT-TH curves is given. The EPT-based aggregate modeling method developed in [5] and [6], and the extensions we introduce, are explained in Section III. Next, the simulation example to illustrate the proposed method is presented in Section IV. The Crolles2 case is discussed in Section V. Finally, Section VI presents our conclusion.

II. CT-TH CURVES

In this section, the CT-TH curve is briefly explained. Two popular approaches to calculate the CT-TH curve are summarized.

A. Working Principle and Purpose of CT-TH Curves

An example of a CT-TH curve is given in Fig. 1. The x-axis denotes the throughput ratio $\delta/\delta_{\max}$, with δ the actual throughput of the system and $\delta_{\max}$ the maximum throughput of the system; the ratio $\delta/\delta_{\max}$ relates to the “utilization” of the integrated processing equipment. The y-axis denotes the mean cycle time ϕ .

For low throughput levels, the cycle time approaches the mean process time t_e of a lot. For increasing throughput, lots experience queueing and the mean cycle time increases. If $\delta/\delta_{\max}$ approaches 1, the system reaches its maximum capacity and the mean cycle time goes to infinity. This vertical asymptote is referred to as the 100% utilization asymptote.

Fig. 1. Example of a CT-TH curve.

The CT-TH curve is used in production planning to make a tradeoff between the throughput of the system and the mean cycle time of the processed lots. On the one hand, a high throughput is desired to ensure high productivity of the capital intensive equipment used. On the other hand, the mean cycle time should be limited to ensure good cycle time performance.

B. Simulation Approaches to Generate CT-TH Curves

Simulation approaches to calculate the CT-TH curve typically use a simulation model that includes several factory-floor realities, such as machine down and repair, setup, operator behavior, and product mix. CT-TH curves may then be derived from the simulation model by fitting a curve to the simulation input and output data. Examples of CT-TH curve fitting approaches can be found in [8]–[10].

Simulation approaches allow the inclusion of many details of the factory floor to arrive at an accurate CT-TH curve prediction. However, simulation model evaluations are often computationally expensive. Furthermore, for each modeled aspect, data has to be obtained from the factory floor. Thus, detailed simulation modeling may involve considerable development and maintenance time.

C. Analytical Approaches to Calculate CT-TH Curves

A popular analytical approach to obtain the CT-TH curve is to use a closed-form $G/G/m$ queueing expression such as (1) (see [2], [3]). Such an analytical approach is insightful and computationally cheap. Relation (1) originates from Kingman's $G/G/1$ approximation [11]. Sakasegawa [1] extended Kingman's equation for workstations with multiple parallel servers under the condition that inter-arrival and process times are exponentially distributed (the $M/M/m$ queue). Whitt [2] proposed (1) for the $G/G/m$ queue. Appropriate values for t_e and c_e^2 may be calculated by including the contribution of the outages in the process time distribution, following [3]. Alternatively, an approach similar to Jacobs *et al.* [7] may be followed, in which t_e and c_e^2 are calculated directly from measured arrival and departure events.

$G/G/m$ queueing approximation (1) assumes that when a machine is processing, its production rate does not depend

on the workload. For integrated processing type of equipment with multiple lots in process at the same time, this is typically not true. Morrison and Martin [12] account for the multi-processing behavior by including an additional delay. However, they note that interactions among the various tool chambers are ignored in their method and that variation in process time and non-ideal tool availability will decrease the accuracy of the predictions. Connors *et al.* [13] developed several analytical queueing expressions for various tools encountered in semiconductor manufacturing, including conveyor tools and multi-sequence batch tools. Various details such as setup, load, and unload times as well as incapacitation are modeled explicitly. This approach requires, like in detailed simulation modeling, that for each modeled aspect data has to be obtained from the factory floor.

III. EPT-BASED AGGREGATE MODELING METHOD

Recently, a new EPT-based aggregate modeling method has been proposed by Kock *et al.* [5], [6], which accounts for workload dependence without the need to model the integrated processing machine in detail. In this section, we describe the concept of the method as well as the approach we introduce to obtain data on arrivals and departures from the semiconductor factory floor, and to deal with the limited number of arrivals and departures collected in the measurement period.

A. Model Concept

The EPT-based aggregate modeling method proposed in [5] and [6] approximates a workstation by a multi-server station similar to the $G/G/m$ system, the main difference being that the mean and variance of the process time distribution depend on the number of lots in the system. Fig. 2 shows an example of a workstation with three integrated processing tools (a), and the corresponding aggregate model (b). The number of servers in the aggregate model, m , we choose equal to the number of machines in the workstation. In the aggregate model, lots arrive according to some arrival process in an infinite, first-in-first-out buffer that feeds the parallel servers. After completion of service, lots leave the system (no blocking after service). In the aggregate model, service starts if a machine is or becomes idle and lots are present in the queue (non-idling assumption). The process time of a lot on an aggregate server is denoted as the EPT and is sampled from an EPT distribution at the moment it starts processing according to the aggregate model. The mean and coefficient of variation (CV) of the EPT distribution depend on the momentary number of lots in the system.

Let w be the number of lots present in the aggregate system at the EPT start of lot i (including lot i itself). For each workload level, a bucket j is defined. If there are w lots in the system upon process start of lot i , then the process time of lot i is sampled from a distribution with distribution parameter values corresponding to bucket $j = w$. Following [5] and [6], we use an independent process time distribution for each bucket. To limit the number of buckets in the aggregate model, we define a highest bucket number N . For $w > N$ the process time is sampled from bucket N . We typically set N equal to

Fig. 2. (a) Example of a workstation. (b) Corresponding structure of the aggregate model.

the workload level for which the maximum production rate has been reached.

In the aggregate model, two different cases can be identified; an EPT starts upon arrival of a new lot when server capacity in the aggregate model is still available, or a new EPT starts upon departure of a lot when lots are waiting to be processed at one of the servers in the aggregate model. Accordingly, we define buckets $j \in [1, \dots, m]$ to sample EPTs starting upon arrival, and buckets $j \in [m, \dots, N]$ to sample EPTs starting upon departure. We denote the mean EPT and CV in each bucket by $t_{e,j}^a$ and $c_{e,j}^a$ for EPTs sampled upon arrival, and $t_{e,j}^d$ and $c_{e,j}^d$ for EPTs sampled upon departure. Note that Kock *et al.* [5], [6] do not distinguish between EPTs starting upon arrival and EPTs starting upon departure in their aggregate model. Distinguishing these two types of EPTs is a necessary ingredient in the curve fitting approach we will introduce in Section III-D, because the EPT distribution parameters depend differently on the workload for the two cases.

B. Obtaining Arrivals and Departures

The workload-dependent distribution parameters for the aggregate model are estimated from lot arrival and departure times registered at the workstation under consideration. Kock *et al.* [5], [6] tested their method in a simulation environment, where lot arrivals and departures can be easily obtained from the simulation output. In a semiconductor factory, arrivals and departures can usually be obtained from the MES. However, this typically requires some processing and filtering of data stored in the MES.

The MES usually registers several lot attributes, including the lot status (e.g., waiting for processing or processing), the process step, and the identity of the machine where the lot is processed. Each time an attribute of a lot in the factory changes, the MES stores the corresponding values and adds a time stamp.

Next we explain how lot arrivals and departures can be derived from such MES data and also give some examples of exceptions to be accounted for. Most lots that arrive at a workstation first have to wait in a queue, 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 usual situation occur. One exception occurs when a lot temporarily gets the status “on hold” while it is in the queue; 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 example 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.

C. Calculation of EPT Realizations

We format the arrivals and departures obtained from the MES as a list, to which we refer to as the event list. Each event in the list contains the lot ID, the event type (arrival or departure) and the time of occurrence of the event. The event list is sorted on increasing time order. We then use a slightly modified version of the EPT algorithm of Kock *et al.* [5], [6] to calculate the EPT realizations. Our modification is that we distinguish between EPTs that start upon arrival of a lot, and EPTs that start upon departure of a lot. Our algorithm is described in detail in Appendix A.

The EPT algorithm calculates EPT realizations from the event list and assigns each EPT to a bucket (corresponding to a workload level). The EPT algorithm takes the aggregate model viewpoint to reconstruct the EPT realizations from the measured arrivals and departures. A new EPT starts in one of the following two cases.

- 1) A lot arrives while less than m lots are present in the (aggregate) system; because at least one of the servers in the aggregate model is idle, the EPT realization of this lot starts immediately (upon arrival).
- 2) A lot departs while leaving $n \geq m$ lots behind; now, one server becomes idle, and is immediately filled with a new lot. This implies an EPT start for the new lot (upon departure of the previous lot).

An EPT ends when a lot departs. Upon the departure of lot i , the algorithm calculates the EPT by subtracting the EPT start of lot i from the departure time of lot i . The EPT is assigned to bucket w , with w the number of lots in the system upon the EPT start. In case a departing lot i has overtaken m or more lots, its EPT has not yet started according to the aggregate model; already m EPTs were running when lot i arrived. In that case, the EPT algorithm picks one of the available EPT starts of other lots. Three rules have been investigated in [5] and [6] to pick an EPT start. In this paper, an EPT start is randomly chosen from the available starts. The chosen EPT start is immediately re-started, because the lot from which the EPT start has been used has not yet departed.

As an example, Fig. 3(a) shows the lot-time diagram of four lots that are processed in first-in-first-out order. For $m = 1$ and $m = 2$ the EPT realizations calculated by the EPT algorithm are visualized in the grey box. In Fig. 3(b), the lots overtake each other (last-in-first-out). Again the EPT realizations are given, but note that for $m = 2$ in Fig. 3(b), other EPTs are possible because the EPT start is randomly chosen for Lots 3 and 4. Also note that some EPTs start upon arrival of a lot

Fig. 3. Lot-time diagrams of arrivals and departures and the calculated EPT realizations (with bucket j indicated between parenthesis). (a) Without overtaking. (b) With overtaking.

(e.g., the EPT started by Lot 1 in Fig. 3(a)), whereas other EPTs start upon departure of a lot (e.g., the EPT started upon departure of Lot 1 in Fig. 3(a)).

Given a set of lot arrivals and departures in a certain measurement period, some lots may have arrived or departed outside this time period. Thus lots may be missing either an arrival or a departure event. For lots that arrived before the start of the measurement period, we set the arrival times equal to the start time of the measurement period. This enables us to correctly determine the number of lots in the workstation at the start of the data collection period. We discard the EPT realizations that start upon reset arrival times. We also discard EPTs that do not end within the collection period.

D. Estimating EPT Distributions

In practice, it may not be possible to accurately estimate the mean and the CV of all N EPT distributions, where N is the workload level for which the production rate has approximately reached its maximum value. For a typical workstation in a wafer fab, N may well be 50 or higher, while the number of collected arrivals and departures is limited to some ten thousands. As a consequence, for the low and high workload levels usually few EPT realizations will occur, which results in inaccurate estimates of t_e and c_e in the low and high buckets. The accuracy of the aggregate model is very sensitive to the estimation of t_e in the highest buckets, because it determines the predicted maximum throughput of the system.

To deal with the limited amount of data, we introduce a curve fitting procedure. We define empirical closed-form approximations and fit these expressions to the t_e and c_e measured for the various workload levels. We fit separate expressions for the EPTs that started upon arrival and the EPTs that started upon departure (see Section III-A). This yields four approximation equations as function of bucket j : $\hat{t}_e^a(j)$, $\hat{c}_e^a(j)$, $j \in [1, \dots, m]$, and $\hat{t}_e^d(j)$ and $\hat{c}_e^d(j)$, $j \in [m, \dots, N]$. We use the following empirical curves: the linear function, the exponential function, and the sum of two exponential functions, with two, three, and six fitting parameters, respectively. Least-squares fitting is used to estimate the parameters of the curves.

E. CT-TH Curve Prediction

We implement the m -server aggregate model as a discrete-event simulation model. Following [5] and [6], we use gamma

distributions to represent the workload-dependent EPT distribution. For lots that start processing in the aggregate model upon arrival, we use an independent gamma distribution for each bucket j , with mean $\hat{t}_e^a(j)$, and CV $\hat{c}_e^a(j)$, $j \in [1, \dots, m]$. For lots that start processing upon departure of a previous lot, we use a gamma distribution for each j with mean $\hat{t}_e^d(j)$, and CV $\hat{c}_e^d(j)$, $j \in [m, \dots, N]$. Gamma distributions are relatively easy to construct because the scale and shape parameters can be readily obtained from the measured EPTs. However, other distributions might have been used.

The CT-TH curve is obtained by simulating the aggregate model for various throughput levels. Note that the aggregate model's parameters have been estimated from arrivals and departures collected at a single throughput level, i.e., the throughput during the data collection period.

IV. ILLUSTRATION

We consider a simulation case representing a workstation with integrated processing machines. The case illustrates the aggregate modeling method described in Section III. Furthermore, we investigate the accuracy of the aggregate model in predicting the mean cycle time.

A. Description of the Case

The test case system is depicted in Fig. 4(a). Lots arrive at the infinite first-in-first-out buffer according to a Poisson process. The workstation consists of four parallel machines M_1 , M_2 , M_3 , and M_4 . Each machine (denoted by the dashed lines in Fig. 4(a)) consists of three sequential process steps, with a one-place buffer between the first and second processes. The first and third process steps may be seen as a preparation and a finalizing process step, respectively (e.g., processes in the track of a litho-cell). The first and third process steps have gamma distributed process times with mean 1.0 and CV 0.5. The second process, which is the bottleneck process in the machine (e.g., the scanner in a litho-cell), has exponentially distributed process times with mean 2.0. A machine M_k is considered to be available for processing if the first process step is idle. The buffer sends a lot to the available machine with the lowest index k .

We model the workstation by the aggregate system depicted in Fig. 4(b), with the number of parallel servers equal to the number of machines in the workstation, i.e., $m = 4$. We implement both the detailed workstation, and the aggregate model in the simulation language χ [14].

B. Obtaining Data on Arrivals and Departures

We simulate 50 000 processed lots in the simulation representation of the workstation at a throughput ratio $\delta/\delta_{\max}$ of 0.8. Arrivals and departures are obtained from the simulation run.

C. Calculation of EPT Realizations

The algorithm in Appendix A is used to calculate EPT realizations, which are assigned to buckets as explained in Section III. The left side of Fig. 5 gives the measured $t_{e,j}^a$

Fig. 4. (a) Simulation case. (b) Corresponding aggregate model.

Fig. 5. Measured EPT distribution parameters and fits for the simulation case.

and c_e^a , whereas its right side plots the measured $t_{e,j}^d$ and $c_{e,j}^d$ (solid black lines). The dotted lines in Fig. 5 show the 95% confidence intervals. We set maximum bucket number N to 30, because for higher buckets $t_{e,j}^d$ and $c_{e,j}^d$ do not significantly change anymore. Note that for very low, and very high values of j , the mean EPTs and CVs cannot be accurately estimated. Hence, we fit approximation functions to the measured mean and CV of the EPT.

D. Estimating EPT Distributions

The fitted curves $\hat{t}_e^a(j)$, $\hat{t}_e^d(j)$, $\hat{c}_e^a(j)$, and $\hat{c}_e^d(j)$ are represented by the solid grey lines. Relations $\hat{t}_e^a(j)$, $\hat{t}_e^d(j)$, $\hat{c}_e^a(j)$, and $\hat{c}_e^d(j)$ are obtained by fitting an appropriate curve to the measured data, using least-squares curve fitting, weighting the measured EPT mean and CV with the number of EPT realizations in bucket j .

Fig. 5 shows that $t_{e,j}^a$ increases for increasing j . Recall that $t_{e,j}^a$ is the mean EPT of lots that immediately start an EPT upon arrival, which equals the time that the lot spends in the system. For $j = 1$, the mean EPT is approximately equal to the sum of the process times of the three sequential process steps in the machine, which is 4.0. For increasing j , the probability that a lot is delayed by other lots already in the machine increases, so $t_{e,j}^a$ increases. We approximate the increasing $t_{e,j}^a$ by the linear function as follows:

$$\hat{t}_e^a(j) = \alpha + \beta(j - 1) \quad j \in [1, \dots, m] \quad (2)$$

where α is the value of $\hat{t}_e^a(j)$ at bucket 1, and β the slope of the linear function. Values α and β obtained by least-squares fitting are equal to 4.24 and 0.23, respectively.

Fig. 5 further shows that $c_{e,j}^a$ is nearly constant. Accordingly, we approximate $c_{e,j}^a$ by a constant function equal to the weighted mean of the measured coefficients of variation in buckets $j \in [1, \dots, 4]$, which is 0.52.

Fig. 5 also shows that $t_{e,j}^d$ decreases for increasing j . Recall that $t_{e,j}^d$ is the mean of EPTs that start upon departure of a previous lot. Hence, $t_{e,j}^d$ can also be seen as the mean interdeparture time per server in the aggregate model. The mean interdeparture time per modeled server decreases for increasing j , because more lots are available for processing on the integrated machines in the workstation, so their production rate increases. The mean interdeparture time decreases until the maximum capacity of the workstation is reached. The maximum capacity is reached when all 12 processes and 4 buffers in the four machines are occupied. We approximate this behavior by the exponential function as follows:

$$\hat{t}_e^d(j) = \theta + (\eta - \theta)e^{-\lambda(j-m)} \quad j \in [m, \dots, N]. \quad (3)$$

In this equation, θ represents the value of $\hat{t}_e^d(j)$ at bucket ∞ . Variable η represents the value of $\hat{t}_e^d(j)$ at bucket m . Variable λ represents the “decay constant” of the exponential curve. We estimate variables θ , η , and λ using a non-linear least-squares fitting procedure; this gives $\theta = 4.07$, $\eta = 2.20$, and $\lambda = 0.32$.

Finally, Fig. 5 shows that $c_{e,j}^d$ increases for increasing j . The cause is that for increasing j , the machine bottleneck process is utilized more and has more influence on the departure process of the machine. Hence, the CV of the interdeparture time approaches the process time CV of the bottleneck process (which is 1.0) for increasing j . We also approximate this behavior by exponential function (3). Non-linear least-squares fitting yields the following parameter values $\theta = 0.61$, $\eta = 0.89$, and $\lambda = 0.39$.

E. CT-TH Curve Prediction

The detailed simulation model of the workstation considered is used to calculate the “real” CT-TH curve of the workstation for throughput ratios $\delta/\delta_{\max}$ ranging between 0.0 and 1.0. For each utilization level, 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.¹

The aggregate model in Fig. 4(b) with $m = 4$ is used to predict the CT-TH curve of the workstation. The fitted expressions for $\hat{t}_e^a(j)$, $\hat{t}_e^d(j)$, $\hat{c}_e^a(j)$, and $\hat{c}_e^d(j)$ are used as the aggregate model’s parameters. We predict the mean cycle time at the same throughput levels for which we calculated the real cycle time, using again 30 replications, a simulation length of 10^5 lots, and a start-up period of $2 \cdot 10^4$ lots.

Fig. 6 shows the real CT-TH curve of the workstation (solid), and the CT-TH curve predicted by the aggregate model (dashed). The dots represent simulation evaluations.

¹We note that we have also experimented with discarding the first $2 \cdot 10^4$ lots in the generation of the arrivals and departures in Section IV-B. However, we observed that this has very little influence on the EPT curve fits obtained.

Fig. 6. CT-TH curve of the simulation test case system and CT-TH curve generated by the proposed EPT method.

The grey lines are the CT-TH curves calculated by $G/G/m$ approximation (1) (with m again the number of machines in the workstation). To determine mean process time t_e and CV of the process time c_e in (1), we assign all measured EPT realizations to a single bucket and calculate the mean and CV.

Fig. 6 clearly shows that the proposed aggregate modeling method estimates the 100% utilization asymptote far more accurately than the $G/G/m$ approximation. Eq. (1) provides an incorrect representation for the multi-processing workstation, while the proposed aggregate model accurately predicts the CT-TH curve up to a throughput ratio of 0.95. Recall that the aggregate model was trained at $\delta/\delta_{\max} = 0.8$. An error in the mean cycle time prediction arises only when the throughput level becomes much higher than the throughput at the training point.

F. Predicting the CT-TH Curve of a Five-Machine Workstation

Finally, suppose that one additional machine is installed. We now predict the CT-TH curve of the five-machine workstation by adding a fifth server to the aggregate model, while still using fitted curves $\hat{t}_e^a(j)$, $\hat{c}_e^a(j)$, $\hat{t}_e^d(j)$, and $\hat{c}_e^d(j)$ measured at the original four-machine workstation. We modify the fitted curves by using $m = 5$ in the expressions used for the curve fits. Furthermore, we decrease the parameter value λ in the exponential function used for $\hat{t}_e^d(j)$ (3) such that the value of j for which $\hat{t}_e^d(j)$ reaches its minimum value increases with 25%. We do this to account for the 25% increase in workstation capacity. We also decrease λ for $\hat{c}_e^d(j)$ to increase the value of j for which $\hat{c}_e^d(j)$ reaches its maximum value with 25%. The aggregate model, with $m = 5$ and the modified curve fits is used to predict the CT-TH curve of the five-machine workstation. The predicted CT-TH curves are depicted in Fig. 7. Fig. 7 shows that the CT-TH curve is again accurately predicted up to a throughput ratio of 0.95. So the aggregate model can be easily adjusted to predict the effect of a change of the number of installed machines.

V. CROLLES2 CASE

CT-TH curves for four workstations in the Crolles2 wafer fab are determined using the EPT-based aggregate modeling

Fig. 7. CT-TH curve of five-machine workstation and CT-TH curve generated by the EPT method using curve fits measured from the four-machine workstation.

method. Crolles2 is a multi-product 300 mm fab in which both high volume products as well as small series and prototype products are produced. Standard production lots contain 25 wafers. Lots are processed in several so-called areas: lithography, implant, etch, thermal treatment, metal, dielectrics, chemical mechanical polishing, wet processing, and metrology. In this section, we first describe the Crolles2 workstations for which the CT-TH curves are calculated. Subsequently, we determine EPT distributions from measured arrivals and departures. Finally, the CT-TH curves are calculated.

A. Modeled Workstations

CT-TH curves are calculated for the Lithography, metal, implant, and critical dimension (CD) measurement workstation. These are all workstations with multiple integrated processing machines.

The lithography workstation consists of 14 track-scanner cells of different types. Lots are manually loaded onto one of the loadports of the machine; next, 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 are developed and hard-baked. When all wafers of a lot have been loaded, the track starts loading the wafers of the next lot if available on the loadport. A track-scanner has four loadports; thus wafers of at most four lots can be processed at the same time, depending on the number of wafers per lot.

The metal workstation consists of all tools that deposit layers of metal on the wafer (i.e., aluminium, copper, and nickel). In Crolles2, the metal workstation consists of 16 machines. A metal tool sequentially loads wafers of up to three lots by means of three available loadports. Wafers pass multiple chambers in the metal tool: vacuum, cleaning, and metal deposit chambers.

The implant workstation includes tools that selectively deposit dopant ions in the surface of wafers. The implant workstation consists of 11 implant tools. From four loadports, wafers of up to two lots are sequentially loaded into one of two

Fig. 8. Measured EPT distribution parameters and fits for the CD measurement workstation.

vacuum chambers. Subsequently, wafers are implanted one by one in the ion implant chamber.

The CD measurement workstation includes tools that measure critical dimensions of wafer patterns created by upstream processes, for process control purposes. Typically, two to four wafers per lot are measured. From three loadports, lots are sequentially loaded into the machine. One by one, wafers pass one of two vacuum chambers and the measurement chamber. Up to three wafers that belong to at most two lots can be in the machine at the same time.

B. Obtaining Data on Arrivals and Departures

At the Crolles2 site, arrivals and departures of about 30 000 to 60 000 processed lots per workstation were obtained from the MES (see Section III-B).

C. Calculation of EPT Realizations

We use the EPT algorithm of Appendix A to calculate EPT realizations for each workstation. The EPT realizations are assigned to buckets, and $t_{e,j}^a$, $t_{e,j}^d$, $c_{e,j}^a$, and $c_{e,j}^d$ are obtained as explained in Section III. We choose the number of parallel servers m in our aggregate models equal to the number of machines in the workstation. N is taken equal to the workload for which the minimum value of $t_{e,j}^d$ has approximately been reached, except when $t_{e,j}^d$ still decreases at the highest measured bucket (see Fig. 8). In that case, we choose N equal to the highest measured bucket.

The left side of Fig. 8 shows the measured $t_{e,j}^a$ and $c_{e,j}^a$ of the CD measurement workstation. The right side of Fig. 8 shows the measured $t_{e,j}^d$ and $c_{e,j}^d$. We have set $N = 52$, which is equal to the highest measured bucket. For reasons of confidentiality, no values on the y-axes are given. The dotted lines in Fig. 8 show the 95% confidence intervals.

D. Estimating EPT Distributions

The fitted curves $\hat{t}_e^a(j)$, $\hat{t}_e^d(j)$, $\hat{c}_e^a(j)$, and $\hat{c}_e^d(j)$ are represented by the solid grey lines in Fig. 8. For $\hat{t}_e^a(j)$, $\hat{t}_e^d(j)$, and $\hat{c}_e^a(j)$, we

Fig. 9. CT-TH curve of the lithography workstation.

Fig. 11. CT-TH curve of the implant workstation.

Fig. 10. CT-TH curve of the metal workstation.

Fig. 12. CT-TH curve of the CD measurement workstation.

use the same functions as in Section IV; we approximate $t_{e,j}^a$ by a linear function, $t_{e,j}^d$ by an exponential function, and $c_{e,j}^a$ by a constant. Fig. 8 shows that $t_{e,j}^d$ still steadily decreases until fairly high workload levels, although wafers of at most two lots can be in process at the same time. This is probably because less wafers per lot are measured at the CD measurement workstation if the queue length is very long, causing the mean EPT of a lot to decrease.

For $c_{e,j}^d$ of the CD measurement workstation, different behavior is observed compared to the simulation test case presented in Section IV. $c_{e,j}^d$ first increases for increasing bucket j (up to bucket 15), and then decreases. To model this behavior, we use a sum of two exponential functions as follows:

$$\hat{c}_e^d(j) = \theta_1 + (\eta_1 - \theta_1)e^{-\lambda_1(j-m)} + \theta_2 + (\eta_2 - \theta_2)e^{-\lambda_2(j-m)}. \quad (4)$$

Herein, $\theta_1 + \theta_2$ represents the value of $\hat{c}_e^d(j)$ at bucket ∞ . Furthermore, $\eta_1 + \eta_2$ represents the value of $\hat{c}_e^d(j)$ at bucket m . Parameters λ_1 and λ_2 represent the decay constants of the first and second exponential curves, respectively. For the first exponential curve, we impose $\lambda > 0$, and $\theta_1 < \eta_1$. Hence, the

first exponential decreases. For the second exponential curve, we impose $\lambda_2 > 0$, and $\theta_2 > \eta_2$. Thus, the second exponential function increases. By imposing $\lambda_2 > \lambda_1$ we accomplish that $\hat{c}_e^d(j)$ first increases, and then decreases.

For the other workstations (lithography, metal, and implant) the same procedure has been followed. For all workstations the curve fits $\hat{t}_e^a(j)$, $\hat{t}_e^d(j)$, $\hat{c}_e^a(j)$, and $\hat{c}_e^d(j)$ have been obtained by least-squares fitting, the measured t_e and c_e values for each bucket j being weighted according to the number of EPT realizations in bucket j . Because of confidentiality, no table with parameter values is given here.

E. CT-TH Curve Prediction

CT-TH curves of the lithography, metal, implant, and CD-measurement workstations are generated using a discrete event simulation implementation of the aggregate models with m set equal to the number of machines in the workstations. We implemented the aggregate models in the language χ . Simulation experiments are carried out for a range of throughput levels, using again 30 replications, and a simulation run length of 10^5 lots, discarding the first $2 \cdot 10^4$ lots.

```

 $n := 0; rs := []; ws := []$ 
loop
  read  $id, ev, \tau$ 
  if  $ev = 'A'$  then
     $n := n + 1$ 
    if  $n \leq m$  then (a1)
       $rs := \text{append}(rs, (id, \tau, n, ev))$  (a2)
    elseif  $n > m$  then
       $ws := \text{append}(ws, id)$ 
    endif
  elseif  $ev = 'D'$  then
     $n := n - 1$ 
    if  $n < m$  then (d1)
       $(t, b, sev) := \text{get}(rs, id)$ 
       $rs := \text{remove}(rs, id)$ 
    elseif  $n \geq m$  and  $id \in rs$  then (d2)
       $(t, b, sev) := \text{get}(rs, id)$ 
       $rs := \text{remove}(rs, id)$ 
       $jd := \text{head}(ws); ws := \text{tail}(ws)$ 
       $rs := \text{append}(rs, (jd, \tau, n, ev))$ 
    elseif  $n \geq m$  and  $id \notin rs$  then (d3)
       $(jd, t, b, sev) := \text{find}(rs, \text{rule})$ 
       $rs := \text{remove}(rs, jd)$ 
       $rs := \text{append}(rs, (jd, \tau, n, ev))$ 
       $ws := \text{remove}(ws, id)$ 
    endif
  write  $\tau - t, b, sev$ 
endif
endloop

```

Fig. 13. EPT algorithm.

Figs. 9–12 present the calculated CT-TH curves in dimensionless form (because of confidentiality). The dashed lines are the predictions generated using the aggregate model (the dots being the simulation evaluations). The solid grey lines are the CT-TH curves calculated by $G/G/m$ approximation (1) (with m again equal to the number of machines in the workstation considered). The mean and variance of the process time distribution in the $G/G/m$ approximation are obtained by assigning all measured EPT realizations to a single bucket (see also Section IV). The x-axis denotes the ratio of throughput δ and throughput δ^* at the working point. The working point is the point on the curve where the workstation was operating during the period of data collection. The y-axis represents the ratio between estimated mean cycle time $\hat{\phi}$ and the mean cycle time at the working point $\hat{\phi}^*$. This implies that the working point is at (1, 1).

The aggregate models accurately estimate the mean cycle time at the working point. The ratio $\hat{\phi}/\hat{\phi}^*$ at the working point of the lithography, metal, implant and CD measurement workstation is 0.943, 0.995, 0.915, and 1.065, respectively. The prediction errors may be caused by the randomly chosen EPT start time if a lot overtakes more than m lots in the considered system. In practice, this may happen due to dispatching strategies, machines with different speeds, and so on. These effects are not modeled explicitly but aggregated in the EPT distribution.

We can only verify the accuracy of the estimated CT-TH curve at the operating point. The simulation case presented in Section IV incorporates similar workload-dependent behavior as the Crolles2 workstations. Based on the results obtained for this simulation case, we expect that the aggregate model predictions for Crolles2 workstations are accurate in the vicinity of the training point. The figures do show that the aggregate models estimate the location of the 100% utilization asymptote (far) more accurately than the $G/G/m$ approximation, again similar to what we found for the simulation cases in Section IV.

VI. CONCLUSION

We have demonstrated that CT-TH curves for workstations consisting of integrated processing tools may be effectively generated using an aggregate $G/G/m$ type of simulation model with workload-dependent process times. The workload-dependent process time distribution parameters are estimated from arrival and departure events measured at the operating workstation. We have introduced a curve fitting procedure to deal with the typically limited number of arrival and departures encountered in semiconductor manufacturing practice. A simulation test case shows that accurate CT-TH curves are obtained based on 50 000 simulated lot arrivals and departures. Also the effect of increasing the installed number of machines on the CT-TH curve is accurately predicted.

As a proof of concept in semiconductor manufacturing, a case from the Crolles2 fab is presented. CT-TH curves for four workstations have been generated using the proposed approach. The arrival and departure data of approximately 30 000–60 000 lots have been extracted from the MES. The generated CT-TH curves are found to be far more accurate than the CT-TH curves generated by means of a well-known closed-form $G/G/m$ queueing expression. This shows that for integrated processing equipment it is crucial to properly take the workload-dependent behavior into account.

The proposed workload-dependent queueing model is very simple. We have implemented it as a discrete-event simulation model, but an analytical queueing model might also be derived. Finally, we note that we have focussed here on the prediction of the mean cycle time. In some cases one may be interested to predict the cycle time distribution, instead of just the mean, e.g., to predict 95% quantiles. The method we have presented here is not necessarily suited for this purpose. This is the subject of our ongoing research.

ACKNOWLEDGMENT

The authors would like to thank A. Kock and I. Adan of the Eindhoven University of Technology, Eindhoven, The Netherlands, and M. Cali, G. Cisinski, V. Gobaille, and B. Lemmen of Crolles2, Crolles, France, for their support.

APPENDIX

The EPT algorithm is depicted in Fig. 13. It is slightly modified compared to the algorithm presented in [5] and [6] to store also the event type upon which an EPT starts (arrival or departure). The EPT algorithm uses the following

variables: n represents the current workload (i.e., number of lots in the system), id the lot ID, ev the event type (either an arrival "A" or a departure "D"), and τ the time the event occurred. For each lot id that is in process, list rs stores the EPT start time, the number of lots in the system upon the EPT start, and the event type upon which the EPT started. List ws contains the id of each lot in the system that has not yet started processing. Variable jd also denotes the lot ID if apart from lot id another lot is considered. Variable t denotes an EPT start time, variable b denotes the bucket number, and variable sev the event upon which the EPT started.

The algorithm uses the functions append, get, remove, head, tail and find, which operate on the lists rs and ws . Function append adds an element to the end of the list, function get reads the element with lot id from the list. Function remove removes the element with id from the list. Function head takes the first element in the list and function tail takes all elements except the first. Finally, function find picks one specific element from the list according to a user-defined rule. In this paper, we pick an EPT start randomly from rs .

The EPT algorithm distinguishes the following five cases:

- (a1) A lot arrives after which $n \leq m$ lots are present in the system (so n includes the arrived lot). Capacity is available so an EPT start of lot id , with start time τ , workload n , and event ev , is added to rs .
- (a2) A lot arrives after which $n > m$ lots are present. All m servers are busy, thus the lot is stored in the list ws .
- (d1) A lot departs, $n < m$ lots remain behind. EPT start time t , bucket b , and the event upon which the EPT started sev are retrieved from rs , after which the lot is removed from rs .
- (d2) A lot departs, $n \geq m$ lots remain behind and id of the departing lot is known in rs : (t, b, sev) is retrieved from rs . Next, id is removed from rs . The first lot waiting in ws is added as a new lot start to rs with time τ , workload n and start event ev .
- (d3) A lot departs, $n \geq m$ lots remain in the system, and id of the departing lot is not known in rs . So lot id departs while it has not started processing according to the aggregate model. Then, using function find, we select an alternative lot jd that has already started an EPT. We compute the EPT realization using the start time of jd . Then, lot jd is restarted and lot id is removed from buffer ws .

REFERENCES

- [1] H. Sakasegawa, "An approximation formula $l_q = \alpha\beta^p(1 - \rho)$," *Ann. Inst. Statist. Math.*, vol. 29, no. 1, pp. 67–75, 1977.
- [2] W. Whitt, "Approximating the $G1/G/m$ queue," *Product. Oper. Manage.*, vol. 2, no. 2, pp. 114–161, 1993.
- [3] W. J. Hopp and M. L. Spearman, *Factory Physics: Foundations of Manufacturing Management*, 3rd ed. New York: IRWIN/McGraw-Hill, 2008.
- [4] S. C. Wood, "Simple performance models for integrated processing tools," *IEEE Trans. Semicond. Manuf.*, vol. 9, no. 3, pp. 320–328, Aug. 1996.
- [5] A. A. A. Kock, "Effective process times for aggregate modeling of manufacturing systems," Ph.D. dissertation, Dept. Mech. Eng., Eindhoven Univ. Technol., Eindhoven, The Netherlands, Jun. 2008.
- [6] A. A. A. Kock, L. F. P. Etman, J. E. Rooda, I. J. B. F. Adan, M. van Vuuren, and A. Wierman, "Aggregate modeling of multi-processing workstations," Eurandom, Eindhoven, The Netherlands, Ext. Rep. 2008-032, Aug. 2008 [Online]. Available: <http://www.eurandom.tue.nl/reports>
- [7] 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.
- [8] S. Park, J. W. Fowler, G. T. Mackulak, J. B. Keats, and W. M. Carlyle, "D-optimal sequential experiments for generating a simulation-based cycle time-throughput curve," *Oper. Res.*, vol. 50, no. 6, pp. 981–990, Nov. 2002.
- [9] F. Yang, B. E. Ankenman, and B. L. Nelson, "Efficient generation of cycle time-throughput curves through simulation and metamodeling," *Naval Res. Logist.*, vol. 54, no. 1, pp. 78–93, Feb. 2007.
- [10] J. W. Fowler, S. Park, G. T. Mackulak, and D. L. Shunk, "Efficient cycle time-throughput curve generation using a fixed sample size procedure," *Int. J. Product. Res.*, vol. 39, no. 12, pp. 2595–2613, Aug. 2001.
- [11] J. F. C. Kingman, "The single server queue in heavy traffic," *Proc. Cambridge Philos. Soc.*, vol. 57, no. 4, pp. 902–904, 1961.
- [12] J. R. Morrison and D. P. Martin, "Practical extensions to cycle time approximations for the $g/g/m$ -queue with applications," *IEEE Trans. Autom. Sci. Eng.*, vol. 4, no. 4, pp. 523–532, Oct. 2007.
- [13] D. P. Connors, G. E. Feigin, and D. D. Yao, "A queueing network model for semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 9, no. 3, pp. 412–427, Aug. 1996.
- [14] A. Hofkamp and J. Rooda, *χ 1.0 Reference Manual*, Syst. Eng. Group, Eindhoven Univ. Technol., Eindhoven, The Netherlands, 2007 [Online]. Available: <http://se.wtb.tue.nl/sewiki/chi>

C. P. L. Veeger is currently a researcher with the Systems Engineering Group, Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. His current research interests include 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, multidisciplinary design optimization, and the effective process time method for performance analysis of manufacturing 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 (M'90) 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.