Generating Cycle Time-Throughput Curves using Effective Process Time based Aggregate Modeling

C.P.L. Veeger\(^1\), L.F.P. Etman\(^1\), J. van Herk\(^2\) and J.E. Rooda\(^1\)

\(^1\) Systems Engineering Group, Department of Mechanical Engineering
Eindhoven University of Technology, the Netherlands, +31-40-2474553

\(^2\) NXP Semiconductors, Nijmegen, The Netherlands
email: {c.p.l.veeger, l.f.p.etman, j.e.rooda}@tue.nl

Abstract—In semiconductor manufacturing, cycle time-throughput (CT-TH) curves are often used to make a trade-off between throughput and mean cycle time of bottleneck workstations. To generate CT-TH curves, detailed simulation models, and analytical queueing approximations are typically used. However, detailed models require much development time and are computationally expensive. Analytical models based on for instance the G/G/m queueing approximation may not generate accurate curves for semiconductor equipment, in particular for equipment that have wafers of multiple lots in process at the same time. In this paper we show that a recently developed aggregate modeling approach based on the effective process time is able to generate more accurate CT-TH curves for equipment with multiple process steps.

Index Terms—cycle time, simulation, CT-TH curves, manufacturing performance, factory dynamics

I. INTRODUCTION

Cycle time-throughput curves are helpful in capacity planning in semiconductor manufacturing. Typically, CT-TH curves are determined for the main contributors to the mean cycle time, such as the litho area. Using the CT-TH curve, the X-factor of a workstation can be controlled by choosing the appropriate utilization level.

To calculate the CT-TH curve, the G/G/m queueing approximation (1), see e.g. [3], [8], [9] is often used in practice.

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

Equation (1) assumes \( m \) identical machines that process one lot at a time.

Many tools in a semiconductor manufacturing environment are integrated processing tools [10]. These tools carry out a series of process steps at various process locations or chambers inside the tool. To maximize utilization, typically multiple loadports are present. As a result the flow of wafers inside the machine may contain wafers of multiple lots at the same time. An example of an integrated processing tool is a track-scanner cell, that may be processing wafers of up to three lots at the same time. For integrated processing tools, the G/G/m approximation may become an inaccurate model representation. As a consequence, the calculated CT-TH curve may be inaccurate.

Recently, an aggregate queueing representation has been developed by [6]. The model representation may be viewed as a G/G/m approximation with a utilization dependent process time distribution. The process time distribution is determined using the EPT concept ([4], [3]). Using simulation, [6] shows that this method can provide accurate aggregate model representations of various manufacturing flow line configurations.

The present paper shows how accurate CT-TH curves for integrated processing tools can be obtained using the new EPT-based aggregate modeling method. We have calculated CT-TH curves for four workstations at the Crolles2 semiconductor manufacturing facility: lithography (track-scanners), metal, implant and critical dimension measurement. For each workstation, arrival and departure times of lots at the workstation were obtained from the manufacturing execution system (MES) covering a period of 10 weeks. From this data, EPT distributions were determined, which are used as input for the aggregate model to calculate the CT-TH curve. These curves are validated by comparing the estimated cycle time to the real cycle time at the working point.

The outline of this paper is as follows. In Section III, an overview of methods to calculate CT-TH curves is presented. The EPT-based aggregate modeling method developed in [6] is explained in Section III. Additionally, Section III discusses how to estimate the EPT-distribution when a limited amount of data is available. Next, we present the Crolles2 case in Section IV. Finally, Section V presents our conclusions.

II. CYCLE TIME-THROUGHPUT CURVES

In this section, the concept and use of the CT-TH curve is discussed. Then, two commonly used methods to calculate the CT-TH curve are introduced: simulation model and analytical methods.

A. Working principle and purpose of the CT-TH curve

An example of a CT-TH curve is given in Figure 1. The x-axis denotes the ratio between throughput \( \delta \) and maximum throughput of the system \( \delta_{\text{max}} \). The y-axis denotes mean cycle time \( \varphi \).
For low throughput levels, the cycle time approaches the mean effective process time $t_e$ of a lot. For increasing throughput, lots experience queueing and the mean cycle time will increase. If $\delta/\delta_{max}$ approaches 1, the system reaches its maximum capacity and the mean cycle time goes to infinity. The throughput level for which the cycle time goes to infinity is referred to as the 100% utilization asymptote.

The CT-TH curve is used in production planning to make a trade-off between the throughput of the system (which is proportional to the utilization) 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 calculate the CT-TH curve

Simulation approaches to calculate the CT-TH curve typically use a model that includes several factory-floor realities, such as machine down and repair, setup, operator behavior, product mix, etc. CT-TH curves may then be derived from the simulation model by fitting the simulation output. Examples of CT-TH curve fitting approaches can be found in [7], [11] and [2].

Cycle time performance is often formulated in terms of the X-factor, which is defined as the mean cycle time of a lot at a workstation divided by its theoretical raw process time. Using simulation, [1] study, among others, the sensitivity of the X-factor to the throughput of the workstation.

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 phenomenon data has to be obtained from the factory floor. As a consequence, detailed simulation modeling may involve considerable development and maintenance time.

C. Analytical approaches to calculate the CT-TH curve

A popular analytical approach to obtain the CT-TH curve is to use the G/G/m queueing approximation (1), see e.g. [3], [9]. This relation originates from Kingman’s G/G/1 approximation [5]. Sakasegawa [8] 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 proposed Equation (1) for the G/G/m queue.

Hopp and Spearman [3] explained the G/G/m equation in terms of the so-called effective process time (EPT). They defined the EPT as ‘the time seen by a lot at a workstation from a logistical point of view’. Basically, the EPT includes not only the raw process time, but also all time losses due to down, setup, and any other source of variability. The mean EPT $t_e$ and coefficient of variation $c_e$ of the EPT are used as input parameters for Equation (1).

[4] used the EPT concept as a starting point to calculate EPT distributions from simple events from the factory floor, such as arrivals and departures of lots on workstations, without the need to measure the contributing factors. [4] use the EPT to estimate the cycle time for single lot machines using the G/G/m approximation.

Analytical models are computationally cheap to evaluate, require little input data and are insightful. However, these models are based on restrictive assumptions. One assumption of the G/G/m approximation is that the machines only process one lot at a time. Many tools in semiconductor manufacturing may have wafers from more than one lot in process at the same time.

III. EPT-BASED AGGREGATE MODELING METHOD

Recently, a new EPT-based aggregate modeling method has been developed by Kock et al. [6]. We utilize this new EPT method to determine CT-TH curves for integrated processing tool workstations in Crolles2. The new EPT method is able to approximate workstations that consist of one or more machines that may process multiple lots at a time.

In this section, we describe the aggregate model concept and explain how we use it to build models of integrated processing tool workstations. Subsequently, we describe how EPT-realizations are calculated in this method. Finally, we discuss how to estimate EPT-distributions.

A. Model concept

The EPT-based aggregate modeling method of [6] approximates a workstation by a multi-server station similar to the G/G/m system, with the difference that the mean and variance of the process time distribution depends on the number of lots in the system. The aggregate multi-server station is visualized in Figure 2. The aggregate station consist of $m$ parallel, identical servers, where $m$ is a user-defined, constant parameter. Lots arrive according to some arrival process in an infinite, first-in-first-out (FIFO) buffer that feeds the parallel servers. It is assumed in the aggregate model that service starts when a machine is, or becomes, idle and lots are present in the queue (non-idling assumption). The process time of a lot is sampled from a distribution at the moment it starts processing according to the aggregate model. Kock et al. [6] use a gamma
distribution, in which the mean and variance is dependent on the momentary number of lots in the system.

Let \( w \) be the number of lots present in the system at the process start of lot \( i \) (including lot \( i \) itself). Then, the process time bucket of lot \( i \) equals \( b_i = w \). [6] uses an independent process time distribution for each bucket. In theory, the number of buckets is infinite. However, above a certain bucket number, the behavior of the buckets stays the same since the system is already operating at its full throughput. To limit the number of buckets, [6] define a highest bucket \( N \) (typically \( N \geq m \)). Bucket \( N \) contains all process times registered with \( N \) or more lots in the system.

The input required for this model consists of one EPT-distribution per bucket, hence \( N \) distributions are required. To determine EPT distributions, arrival and departure data is used. For each lot \( i \) departing from the considered workstation, the departure time \( d_i \) is collected, as well as the corresponding arrival time \( a_i \) of the lot in the buffer of the station. This arrival and departure data is translated into EPT-realizations using an EPT algorithm. From the EPT-realizations, EPT distributions per bucket are calculated.

### B. Calculation of EPT-realizations

The input for the EPT algorithm developed in [6] consists of a list of events. An event contains the lot id, the event type (arrival or departure) and the time of occurrence of the event. The EPT algorithm processes the events in order of time. A new EPT starts in either of the following two cases:

1. A lot arrives while less than \( m \) lots are present in the (aggregate) system: since at least one of the servers in the approximation is idle, the EPT-realization of this lot starts immediately at \( a_i \).
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.

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 \( d_i \). A bucket number \( b_i \) is assigned to the EPT, which is equal to the number of lots in the system \( w \) 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 \( m \) EPT’s 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 [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 at \( d_i \), because the lot from which the EPT start has been used has not yet departed. Figure 3a shows EPT-realizations of four lots that are processed in FIFO order, calculated by the EPT algorithm for \( m = 1 \) and \( m = 2 \). Figure 3b shows EPT-realizations of four overtaking lots. For \( m = 2 \) in Figure 3b, other EPT’s are possible because the EPT start is randomly chosen for lot 3 and lot 4.

### C. Estimating EPT distributions

In practice, it may not be possible to accurately estimate the mean and variance of all \( N \) EPT distributions. For example, for a workstation in a wafer fab, \( N \) may be 50 or more. For the lowest and highest buckets typically little or no EPT-realizations are measured during the period of data collection. Following [6], we assume Gamma distributed EPT distributions with mean \( t_e,j \) and coefficient of variation \( c_{e,j} \), where \( j \) is the bucket. To significantly reduce the number of parameters to be estimated we assume an analytical relation \( t_e = \hat{t}_e(j) \) and \( c_e = \hat{c}_e(j) \) and fit these functions to the measured \( t_{e,j} \) and \( c_{e,j} \) values.

[6] observed that in the EPT-calculation two different regions can be identified: an EPT-realization starts upon arrival of a new lot when server capacity in the aggregate model is still available (which implies \( j \leq m \)), or a new EPT-realization starts upon departure of a lot when lots are waiting to be processed on one of the servers in the aggregate model (which implies \( j \geq m \)). In the former case, the EPT-realization corresponds to the time the lot spends in the workstation; in the latter case the EPT-realization corresponds to the interdeparture time with the previously finished lot.

Accordingly, we define two different functions for \( \hat{t}_e(j) \) and \( \hat{c}_e(j) \). \( t_{e,a} \) and \( c_{e,a} \) are the mean EPT and coefficient of variability of EPT-realizations that started upon arrival of a lot. \( t_{e,d} \) and \( c_{e,d} \) are the mean EPT and coefficient of variability of EPT-realizations that started due to the departure of a lot. To estimate \( t_{e,a} \), \( t_{e,d} \), \( c_{e,a} \) and \( c_{e,d} \), we subdivide the EPT-realizations obtained by the EPT-algorithm in realizations that started upon a lot arrival and realizations that started when a machine became idle.
Expressions for $\hat{t}_{e,a}$, $\hat{c}_{e,a}$, $\hat{t}_{e,d}$ and $\hat{c}_{e,d}$ as a function of the bucket are obtained as follows. We select an appropriate curve to represent each relation. The curves we use in this paper are the linear function and the exponential function, with two respectively three parameters. Least squares curve fitting is used to estimate the parameters of the curves.

In our aggregate simulation mode representation, we distinguish between EPT-realizations that start upon a lot arrival and EPT-realizations that start upon a lot departure. For lots that immediately start processing in the aggregate model upon arrival, the process time is sampled from a distribution with mean $t_{e,a}(j)$ and coefficient of variation $c_{e,a}(j)$. For lots that start processing upon a departure, process distributions with $t_{e,d}(j)$ and $c_{e,d}(j)$ are used. Note again that [6] do not make this subdivision in their aggregate model, and do not build curve fits for $t_{e,a}$ and $c_{e,a}$.

IV. Crolles2 Case

Cycle time-throughput curves are determined for four workstations in the Crolles2 Alliance wafer fab using the EPT-based aggregate modeling method. Crolles2 is a mid-volume multi-process multi-product 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 describe the Crolles2 workstations for which the CT-TH curve is calculated. Subsequently, we elaborate how EPT data (arrivals and departures) are obtained and filtered. Next, we calculate EPT-realizations and estimate EPT distributions. Finally, CT-TH curves are calculated.

A. Considered Workstations

Lithography, metal, implant and critical dimension (CD) measurement workstations are considered. These are all multi-machine workstations, where the tools may be qualified for different subsets of process recipes.

The lithography workstation consists of 14 track-scanner cells of different types. Lots are manually loaded on the machine. Wafers are sequentially loaded onto 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. After 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 consist 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 in one of two 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 in the machine. One by one, wafers pass one of two vacuum chambers and the measurement chamber. Up to three wafers can be in the machine at the same time, that belong to at most two lots.

B. Lot arrivals and departures

Lot arrivals and departures are needed to calculate a workstation’s CT-TH curve using the EPT method. At the Crolles2 site, these events were obtained from the manufacturing execution system (MES).

1) MES data: Table I gives an example of a few lines of raw data extracted from the MES for a particular workstation. Each line contains a lot identification code (LotID), the status of the lot, the start and end time of the status, the name of the process, and the name of the machine on which the process was performed. All values in Table I are fictitious. Table I is sorted on LotID. All lines with the same LotID are sorted on start time. In most cases, a lot arrives at a workstation and first waits in a buffer. Then it is processed after which it departs from the workstation. The first two lines in Table I give an example of such a lot (lot1). In this case, the arrival time of the lot is taken equal to the start time of the ‘wait’ status ($t = 0$). The departure time is taken equal to the stop time of the ‘Running’ status ($t = 15$). The other lines of Table I give examples of lots for which filtering is required.

2) Filter MES data: First, the time that a lot is ‘on hold’ while it is in the workstation’s buffer is removed from the data. The ‘on hold’ status means that the lot becomes temporarily unavailable for processing because of a quality problem. While the lot is on hold in a buffer, it is not available for processing and has no influence on the cycle time of other lots in the

<table>
<thead>
<tr>
<th>LotID</th>
<th>Status</th>
<th>Start</th>
<th>Stop</th>
<th>Process Name</th>
<th>Machine name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lot1</td>
<td>Wait</td>
<td>10</td>
<td>10</td>
<td>ExposeLines</td>
<td>Litho01</td>
</tr>
<tr>
<td>Lot1</td>
<td>Running</td>
<td>15</td>
<td>15</td>
<td>ExposeLines</td>
<td>Litho01</td>
</tr>
<tr>
<td>Lot2</td>
<td>Wait</td>
<td>10</td>
<td>10</td>
<td>ExposeLines</td>
<td>Litho02</td>
</tr>
<tr>
<td>Lot2</td>
<td>Running</td>
<td>15</td>
<td>15</td>
<td>ExposeLines</td>
<td>Litho02</td>
</tr>
<tr>
<td>Lot2</td>
<td>RunningHold</td>
<td>20</td>
<td>20</td>
<td>ExposeLines</td>
<td>Litho02</td>
</tr>
<tr>
<td>Lot2</td>
<td>WaitHold</td>
<td>25</td>
<td>25</td>
<td>ExposeLines</td>
<td>Litho02</td>
</tr>
<tr>
<td>Lot2</td>
<td>Wait</td>
<td>30</td>
<td>30</td>
<td>ExposeLines</td>
<td>Litho02</td>
</tr>
<tr>
<td>Lot2</td>
<td>Running</td>
<td>40</td>
<td>40</td>
<td>ExposeLines</td>
<td>Litho02</td>
</tr>
<tr>
<td>Lot3.0</td>
<td>Wait</td>
<td>10</td>
<td>10</td>
<td>ExposeVias</td>
<td>Litho03</td>
</tr>
<tr>
<td>Lot3.0</td>
<td>Running</td>
<td>10</td>
<td>15</td>
<td>ExposeVias</td>
<td>Litho03</td>
</tr>
<tr>
<td>Lot3.1</td>
<td>Wait</td>
<td>5</td>
<td>10</td>
<td>ExposeVias</td>
<td>Litho03</td>
</tr>
<tr>
<td>Lot5</td>
<td>Wait</td>
<td>0</td>
<td>5</td>
<td>ExposeLines</td>
<td>Litho04</td>
</tr>
</tbody>
</table>
buffer. Therefore, from an EPT point of view, hold time should not be included in the EPT.

Hold time of a lot only adds to the cycle time of the hold lot itself. If holds occur frequently, the extra delay due to holds can be calculated and added to the mean queue time. In Table I, the ‘waithold’ status in line 7 means that the lot is on hold while it is in the buffer. Therefore, this line is removed. Line 5 shows a ‘RunningHold’ status, which means that the lot is on hold while occupying a machine. In this case, the lot does have influence of the cycle time of other lots. As a consequence, this line is not removed from the data.

Another example of an exception that may be encountered is merging of lots. Wafers arrive in different lots but are reunited into one FOUP and processed together. From an EPT point of view, the arrival of the reunited lot is when the last set of wafers arrives. The departure is when the reunited lot has finished processing. This means that data belonging to the first arriving wafer sets can be removed. In the example shown in Table I, wafer sets 3.0 and 3.1 are reunited and depart at the same time. The filter removes the lines of wafer set 3.0, because it arrived first.

Filtering for waithold and lot merging reduces Table I to Table II, from which arrivals and departures can be obtained. 3) Determine arrivals and departures: Arrivals and departures are defined respectively as the begin and end of an uninterrupted time interval in which a lot is (available for) a specific process. In the filtered data, this means that an arrival is the first occurring start time of a group of lines that have the same lotID, process name, and tool name, and that happen in succession of time. A departure is the end of the last occurring running status in such a group of lines. In Table II, this means that two arrivals occur at time 0, one at time 5, and one at time 30. Two departures occur at time 15, one at time 25, and one at time 45. Lot2 was on waithold between time 25 and 30 so the lines belonging to lot2 do not happen in succession of time. Therefore, lot2 generates two arrivals.

4) Post-filtering: Lots that arrive in the data collection period, may depart after this period, and lots that depart in the collection period may arrive before this period. Thus, lots may be missing either an arrival or a departure. For lots that have a departure outside the measurement period, we do not calculate an EPT-realization. For the arrivals that occur before the collection period, we assume that those lots arrived at the start of the measured period. This enables us to determine the number of lots in the workstation at the start of the collection period. The EPT-realizations that originate from lots with the arrival time reset are discarded.

C. Estimating EPT-distributions

We use the EPT-algorithm of [6] to calculate EPT-realizations of the four considered Crolles2 workstations using the arrivals and departures obtained from the MES. From this data we obtain \( t_{e,a} \) and \( c_{e,a} \) and \( c_{e,d} \) (see Section III). We choose the number of parallel servers \( m \) in our aggregate model representation equal to the number of machines in the workstation. \( N \) is set equal to the maximum bucket number for which EPT-realizations were obtained in the measurement period.

The left side of Figure 4 shows the measured \( t_{e,a} \) and \( c_{e,a} \) of the CD measurement workstation. The right side of Figure 4 shows the measured \( t_{e,a} \) and \( c_{e,d} \). For reasons of confidentiality, no values on the y-axes are given. Additionally, Figure 4 shows the 95% confidence interval for \( t_{e,a} \) and \( t_{e,d} \). Furthermore, the fitted relations \( \hat{t}_{e,a}(j), \hat{t}_{e,d}(j), \hat{c}_{e,a}(j) \) and \( \hat{c}_{e,d}(j) \) are visualized.

For \( \hat{t}_{e,a}(j) \) we use the linear function in Equation (2):

\[
\hat{t}_{e,a} = \alpha + \beta(j-1). \quad (2)
\]

Herein, \( \alpha \) is the value of \( \hat{t}_{e,a}(j) \) at bucket 1. We set \( \alpha \) to a user-defined value, since no EPT-realizations were obtained in bucket 1. \( \beta \) represents the mean time a lot spends in the system when it arrives in an empty system. \( \beta \) is the slope of the linear curve, which is obtained by least-squares curve fitting. Herein, the mean EPT for each bucket \( t_{e,a,j} \), \( j = 1, ..., m \) is weighed according to the number of EPT-realizations in bucket \( j \). \( \beta \) will typically be positive. Hence \( t_{e,a} \), representing the mean time a lot spends in the system, typically increases. \( \hat{t}_{e,a}(1) \) is the mean effective process time on a machine when a lot arrives in an empty system. For increasing buckets, a lot arrives when multiple machines are occupied. Lots may not start processing immediately on an idle machine, because machines are typically qualified for a subset of process recipes. Instead, lots may wait for an occupied machine and the time they spend in the workstation increases. The higher the bucket, the more machines will be occupied thus the chance that a lot has to wait increases. As a result, the mean EPT \( t_{e,a} \) will increase.

For \( \hat{t}_{e,d}(j) \), we use the exponential function in Equation (3):

\[
\hat{t}_{e,d}(j) = \gamma + (\delta - \gamma) e^{-\lambda(j-m)}. \quad (3)
\]

Herein, \( \gamma \) represents the value of \( \hat{t}_{e,d}(j) \) at bucket \( \infty \). Variable \( \delta \) represents the value of \( \hat{t}_{e,d}(j) \) at bucket 0. Variable \( \lambda \) represents the ‘decay constant’ of the exponential curve. We estimate variables \( \gamma, \delta, \lambda \) using a non-linear least-squares fitting procedure in which the mean EPT for each bucket \( t_{e,d,j} \), \( j = m, ..., N \) is weighed according to the number of EPT-realizations in bucket \( j \). As explained in Section III, \( \hat{t}_{e,d}(j) \) can be seen as the mean inter departure time of lots per modeled machine in the aggregate model. \( \hat{t}_{e,d}(j) \) decreases for increasing bucket number because CD measurement machines

<table>
<thead>
<tr>
<th>LotID</th>
<th>Status</th>
<th>Start</th>
<th>Stop</th>
<th>Process Name</th>
<th>Machine name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lot1</td>
<td>Wait</td>
<td>0</td>
<td>10</td>
<td>ExposeLines</td>
<td>Litho01</td>
</tr>
<tr>
<td>Lot1</td>
<td>Running</td>
<td>10</td>
<td>15</td>
<td>ExposeLines</td>
<td>Litho01</td>
</tr>
<tr>
<td>Lot2</td>
<td>Wait</td>
<td>0</td>
<td>10</td>
<td>ExposeLines</td>
<td>Litho02</td>
</tr>
<tr>
<td>Lot2</td>
<td>Running</td>
<td>10</td>
<td>15</td>
<td>ExposeLines</td>
<td>Litho02</td>
</tr>
<tr>
<td>Lot2</td>
<td>Running</td>
<td>15</td>
<td>20</td>
<td>ExposeLines</td>
<td>Litho02</td>
</tr>
<tr>
<td>Lot2</td>
<td>Running</td>
<td>20</td>
<td>25</td>
<td>ExposeLines</td>
<td>Litho02</td>
</tr>
<tr>
<td>Lot2</td>
<td>Wait</td>
<td>30</td>
<td>40</td>
<td>ExposeLines</td>
<td>Litho02</td>
</tr>
<tr>
<td>Lot2</td>
<td>Running</td>
<td>40</td>
<td>45</td>
<td>ExposeLines</td>
<td>Litho02</td>
</tr>
<tr>
<td>Lot3.1</td>
<td>Wait</td>
<td>5</td>
<td>10</td>
<td>ExposeVias</td>
<td>Litho03</td>
</tr>
<tr>
<td>Lot3.1</td>
<td>Running</td>
<td>10</td>
<td>15</td>
<td>ExposeVias</td>
<td>Litho03</td>
</tr>
</tbody>
</table>
can process two lots at the same time. So for increasing buckets above $m$, more machine capacity is used and the inter departure time decreases accordingly until the maximum machine capacity is approached.

For $\hat{c}_{e,a}(j)$ and $\hat{c}_{e,d}(j)$, we use a linear curve (Equation (2)) with $\beta$ is zero. We use a least-squares method to estimate $\alpha$ in which again $c_{e,a,j}$, $j = 1, ..., m$ and $c_{e,d,j}$, $j = m, ..., N$ are weighed according to the number of EPT-realizations in bucket $j$. We are currently investigating whether a better fit can be obtained and how it influences the accuracy of the prediction. The current impression is that the slope of the curve fit to $\hat{c}_e$ has a much larger influence than the slope of the curve fit to $\hat{c}_c$.

To obtain $\hat{t}_{e,a}(j)$, $\hat{t}_{e,d}(j)$, $\hat{c}_{e,a}(j)$ and $\hat{c}_{e,d}(j)$ for the metal, lithography and implant workstations we use the same curve fitting approach.

D. CT-TH curves

CT-TH curves of the lithography, metal, implant and CD-measurement workstations are calculated using the aggregate modeling method described in Section III. EPT-distributions are calculated using $t_{e,a}(j)$, $t_{e,d}(j)$, $\hat{c}_{e,a}(j)$ and $\hat{c}_{e,d}(j)$. For aggregate model parameter $m$ we take the number of machines in the workstation.

Figure 5, 6, 7 and 8 show dimensionless (solid) curves estimated by the aggregate modeling method presented in this paper. The dots in the solid curve indicate the calculated values of the mean cycle time. The dimensionless curves predicted by G/G/m approximation (1) is dashed (again, $m$ is the number of machines in the workstation). The mean and variance of the process time distribution in the G/G/m approximation are obtained using the single-lot EPT method of [4]. This method considers only a single bucket with EPT-realizations for which $t_e$ and $c_e$ are calculated. The x-axis denotes denotes the ratio of the throughput $\hat{\varphi}$ and the throughput at the working point $\varphi^*$. The working point is the point on the curve on which the considered workstation was operating during the period of data collection. The y-axis represents the ratio between estimated mean cycle time $\hat{\varphi}$ and the mean cycle time at the working point $\varphi^*$. This implies that the working point is at $(1,1)$.

The EPT-based aggregate modeling method is able to accurately estimate the mean cycle time in the working point. The ratio $\hat{\varphi}/\varphi^*$ at the working point of the lithography, metal, implant and CD measurement workstation is 0.986, 1.002, 0.917 and 1.075 respectively. The difference between $\hat{\varphi}$ and $\varphi^*$ 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 etc. These effects are not modeled explicitly but aggregated through the modified ept start time.

The new EPT method estimates the location of the 100% utilization asymptote more accurately than the G/G/m approximation. The G/G/m approximation incorrectly models that the workstation is 100% utilized when there are $m$ lots in the system because it assumes single lot machines. Integrated processing machine workstations however do not operate at their full throughput when there are $m$ lots in the system. As a result, the G/G/m approximation locates the 100% utilization
asymptote very close to δ∗. This also causes the cycle time in the working point to be overestimated by the G/G/m approximation.

We can only verify the accuracy of the estimated CT-TH curve at the operating point. [6] carried out a simulation study for several flowline configurations, and found that accurate CT-TH predictions (within 10%) are obtained for a wide throughput region around the working point. Only for throughput levels near the maximum capacity prediction errors larger than 10% were found. Still, the location of the 100% utilization asymptote is predicted more accurately than the G/G/m method.

V. Conclusion

In this paper, CT-TH curves are calculated for four Crolles2 workstations using a newly developed EPT based aggregate modeling approach ([6]). The newly developed method is able to model systems that have multiple lots in process at the same time. The new EPT method approximates a workstation by an m-server parallel workstation with a process time distribution that depends on the number of lots in the system. The process time of a lot in the approximation is sampled from a gamma distributed effective process time (EPT) distribution at the moment it starts processing in the approximation. The mean and variance of the EPT distribution from which the process time is sampled depends on the momentary number of lots in the system. The EPT distribution parameters are estimated from lot arrival and departure data obtained from the MES in Crolles2, using the EPT algorithm developed by [6].

The EPT-based aggregate modeling method is able to estimate the mean cycle time in the working point of the considered workstations within 10%. Furthermore, the new EPT-method locates the 100% utilization asymptote more accurately than a G/G/m approximation of the system. The CT-TH curve is considered to be accurate for throughput levels in a region around the working point.

VI. Acknowledgment

We would like to thank Ad Kock and Ivo Adan of Eindhoven University of Technology and Manuel Cali, Gregory Cisinski, Vincent Gobaille and Bart Lemmen of Crolles2 for their support.

REFERENCES