Proceedings of the 2011 Winter Simulation Conference
S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, eds.

AGGREGATE MODELING OF SEMICONDUCTOR EQUIPMENT USING EFFECTIVE PROCESS TIMES

L.F.P. Etman
C.P.L. Veeger
E. Lefeber
I.J.B.F. Adan
J.E. Rooda

Department of Mechanical Engineering
Eindhoven University of Technology
PO Box 513, 5600 MB Eindhoven, THE NETHERLANDS

ABSTRACT

Performance evaluation using queueing models is common practice in semiconductor manufacturing. Analytical closed-form expressions and simulation models are popular in capacity planning and the analysis of equipment configurations. However, the complexity of semiconductor processes complicates the modeling of the equipment. Analytical models lack the required accuracy, whereas simulation models require too many details, making them impractical. Aggregation is a way to overcome this difficulty. The various details are not modeled in detail, but their contribution is lumped in the aggregate model, which makes the model more appropriate for both analysis and simulation.

This paper gives an overview of our efforts to develop a top-down aggregate modeling approach for semiconductor equipment, starting from the effective process time concept inspired by the Factory Physics book of Hopp and Spearman. The strong feature of our modeling approach is that the aggregate model parameters are estimated directly from industrial data (arrival and departure times), without the need to quantify the various details.

1 INTRODUCTION

The development of simple and accurate models for queueing performance analysis in semiconductor manufacturing is an open research problem. With "simple" we mean models that have only few model parameters, and that can be easily used and estimated. With "accurate" we mean models that provide a prediction accuracy that is of practical use in semiconductor applications. This may depend of course on the type of application, but we assume that in most cases an accuracy of 10% or better is desired.

Commonly-used simple models are closed form *G*/*G*/*m* queueing expressions, such as the approximation due to Sakasegawa (1977) and Whitt (1993), which is used in the popular Factory Physics book of Hopp and Spearman (2008). Although useful and insightful, this model lacks the necessary accuracy for the typical equipment encountered on the semiconductor factory floor. Modifications to (partially) account for typical semiconductor equipment characteristics such as the simultaneous processing of wafers of multiple lots have been proposed in e.g., Morrison and Martin (2007b) and Morrison and Martin (2007a).

Alternatively, discrete-event simulation may be used to arrive at a sufficiently accurate representation of the semiconductor equipment. Simulation modeling allows the inclusion of all relevant factory floor details. This obviously requires the collection of all the necessary input data regarding the various model elements. A a result, a detailed simulation model becomes computationally very expensive and requires significant development time. Including too many details makes simulation modeling impractical.

Aggregation is a way to overcome this difficulty. The various details are not modeled in detail, but their contribution is lumped in the aggregate model. For a simulation model this means that fewer details need to be modeled explicitly saving considerable amount of computation and development time. Examples of aggregations used in simulation model building can be found in Brooks and Tobias (2000), Rose (2000), Johnson, Fowler, and Mackulak (2005), and Rose (2007). An analytical result regarding aggregation is the flow equivalent server (FES) due to Norton (1926), see Chandy, Herzog, and Woo (1975). The FES is a single-server representation of a closed queueing network of exponential servers. Also Sakasegawa's and Whitt's closed-form *G*/*G*/*m* queueing approximation may be used as aggregate model. Hopp and Spearman express this approximation in terms of the *effective process time* to account for the contribution of preemptive and non-preemptive outages (e.g., setup, respectively breakdown) to the workstation capacity and variability. Thus the natural process time and the various outages are lumped together into the effective process time.

Yet, the aforementioned aggregation techniques without exception assume that the contributing components are *known* to derive the aggregate model. In practice this is often not possible, since the necessary data are not available or difficult to obtain. This lack of data is also the main obstacle in the development of many detailed simulation models. This motivated us to seek for a top-down instead of bottom-up aggregate modeling approach, which allows to estimate the aggregate model parameters directly from arrival and departure data collected on the factory floor without the need to quantify the contributing details.

Our research on the Effective Process Time as method for top-down aggregate model building was initiated around 1998, inspired by the aforementioned Factory Physics book (Hopp and Spearman 1995). Our first publication was presented in 2001 at the ASMC in Munich (Jacobs, Etman, van Campen, and Rooda 2001). Several publications have followed since. During the last ten years we have developed three generations of aggregate models for equipment in semiconductor manufacturing. In particular we focused on the development of aggregate models for equipment that carry out a series of process steps.

In this paper we give an overview of the three aggregate modeling methods we have developed so far, and explain the ideas behind them. We hope this serves as a source of inspiration for other researchers and practitioners. We start from the Effective Process Time as defined by Hopp and Spearman (1995). Subsequently, we present the three types of aggregate models and give some results of applications to semiconductor equipment. We conclude with presenting the general framework of top-down EPT-based aggregate modeling, and give some suggestions for further research.

2 EPT ACCORDING TO HOPP AND SPEARMAN

Hopp and Spearman (1995) originally defined the Effective Process Time (EPT) as 'the process time seen by a lot at a workstation from a logistical point of view'. The EPT includes besides the natural processing time, also the time during which the lot could have been processed but for some reason was not. Such additional delay may be due to maintenance, setup, or operator unavailability, and has to be accounted for in the calculation of the utilization of the workstation. The effective utilization is typically higher than the utilization calculated from the natural processing time only. What is more, such additional delays may also influence the variability in processing experienced at the workstation. As a result, the variability of the effective process time is typically higher than the theoretical variability. This is very much the case for the highly automated but failure prone and setup-intensive semiconductor equipment.

Hopp and Spearman (2008) use the mean and variance of the EPT in their representation of Sakasegawa's and Whitt's *G*/*G*/*m* approximation of the mean cycle time of an *m*-server workstation with infinite buffer:

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

Herein, t_a and c_a^2 are the mean and the squared coefficient of variation of the inter-arrival time, t_e and c_e^2 the mean and squared coefficient of variation of the effective process time at the station, and u the effective

utilization of the workstation, defined as $u = \frac{t_e}{mt_a}$. Due to the nonlinearity present in this expression, using t_e instead of the natural processing time t_0, yields a significantly higher predicted cycle time value (depending on u). An additional increase in cycle time follows from using c_e^2 instead of the squared coefficient of variation of the natural process time c_0^2. This way of presenting Equation (1) is crucial in the understanding of the cycle time performance of semiconductor equipment (and any workstation for that matter). The aggregated contribution of the natural processing time and the various disturbances is what finally matters for the actual cycle time performance experienced on the factory floor.

Hopp and Spearman (2008) have derived a set of equations to calculate t_e and c_e^2 from data on the natural processing time, preemptive outages and non-pre-emptive outages. They define an outage as the time that a machine is unable to process lots. A pre-emptive outage occurs when a lot is in process, and is temporarily halted, for instance due to machine breakdown. A non-preemptive outage delays the process start of a lot, for instance due to setup. Consider for instance the case of setup. Then the t_e and c_e^2 may be calculated from (Hopp and Spearman 2008):

$$t_e = t_0 + \frac{t_s}{N_s}, \qquad c_e^2 = \frac{\sigma_0^2 + \frac{\sigma_s^2}{N_s} + \frac{N_s - 1}{N_s^2} t_s^2}{t_e^2}, \tag{2}$$

with N_s the average number of lots after which setup takes place, σ_0 the standard deviation of the natural process time, t_s the mean setup time, and σ_s the standard deviation of the setup time. The probability of a setup after processing a lot is assumed to be $1/N_s$ regardless of how many lots have been processed since the previous setup. Similar expressions have been derived for breakdown (non-preemptive outages), assuming exponentially distributed time to failure and generally distributed time to repair, refer to Table 8.2, page 261 in Hopp and Spearman (2008). The equations for breakdown and setup are applied consecutively if multiple outages are present (t_e becomes t_0, and so on).

3 ESTIMATING EPTs FROM ARRIVALS AND DEPARTURES

In practice, data regarding the various outages is only partially available and often difficult to obtain. What is more, the outages may not behave according to the assumptions made above. This may hinder the application of the equations in Table 8.2 of Hopp and Spearman (2008) to calculate the t_e and c_e^2 from the various outages.

The idea that ignited our research was presented in Jacobs, Etman, van Campen, and Rooda (2001), Jacobs, Etman, van Campen, and Rooda (2003): estimate t_e and c_e^2 from arrival and departure events measured at the workstation in operation, without quantifying the various outages. The original m-machine workstation with the various outages is modeled by means of m aggregate servers and an infinite buffer. For each lot processed at the workstation an EPT realization is calculated. These EPT realizations are reconstructed from the arrival and departure times measured at the physical workstation, by pretending that the arrivals and departures happened at the servers of the aggregate model. The collected events are sorted and processed in order of time. An EPT starts if:

- a lot arrives while there are fewer than m lots in the workstation, or
- a lot departs while the number of lots that remain in the workstation is larger than or equal to m.

In both cases, an aggregate server is or has become idle and is available to start processing. These EPT starts are assigned to the respective idle server. An EPT ends if:

- a lot departs from the workstation.

The resulting EPT realization is then equal to the departure time minus the EPT start time at the aggregate server. The various EPT realizations provide an empirical EPT distribution, from which the mean and coefficient of variation can be easily calculated.

The concept is best understood by means of an example: a workstation consisting of a *single* machine that processes lots in first-in-first-out (FIFO) sequence from an infinite buffer, subject to setup, breakdown and operator unavailability. A Gantt chant chart of five processed lots is depicted in Figure 1. The first lot arrives at $t = 0$. It needs a setup of two time units, and is subsequently processed. It is finished at $t = 6$. Meanwhile the second lot has arrived at $t = 4$. Obviously, this lot has to wait until the first lot is completed. Unfortunately at $t = 6$ the operator is not available to do the setup of the new lot. Setup is suspended until $t = 7$. The actual processing starts at $t = 8$ and is completed at $t = 12$. Similar events occur for the third, fourth and fifth lot depicted in Figure 1. Note that from $t = 17$ until $t = 19$ the workstation is empty.

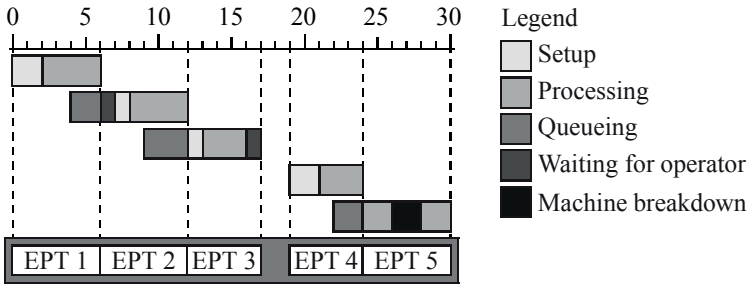


Figure 1: Gantt chart of five lots processed at a single machine workstation.

The EPTs are reconstructed as follows. Upon arrival of the first lot the workstation is empty so the EPT starts. This EPT ends when the first lot departs. A new EPT is started immediately since the second lot has already arrived which means the number of lots remaining in the workstation is equal to the number of machines in the workstation $m = 1$. This second EPT realization ends at $t = 12$. Similarly we obtain a third EPT realization for the third lot. Then, at $t = 17$, the system is empty, which implies that no new EPT is started until the fourth lot arrives at $t = 19$. The EPT realization of the fifth lot follows suit. The calculated EPTs are depicted in Figure 1.

In the example, the lots are processed in FIFO sequence. Then, the EPT realizations may also be obtained using the following equation:

$$\mathrm{EPT}_i = d_i - \max(a_i, d_{i-1}) \tag{3}$$

where EPT_i is the EPT realization due to the i^{th} lot departing from the machine, d_i is the departure time of the i^{th} departing lot, a_i is the arrival time of the i^{th} departing lot, and d_{i-1} is the departure time of the $(i-1)^{\text{th}}$ departing lot. Essentially, this is an inverse use of the well-known sample path equation. Instead of computing a series of departure events from the sample path equation, we compute EPT realizations.

Also for multi-machine workstations Equation (3) may be used to calculate the EPTs, provided the (effective) processing of lots at the (physical) workstation was carried out in FIFO sequence. Equation (3) is simply applied for each machine separately, with d_i the departure time of the i^{th} lot departing from that machine, and a_i the arrival time of the i^{th} lot departing from that machine. Gathering the EPTs from the machines provides the workstation EPT distribution. A nice property is that the mean and variance are utilization independent; it is valid to use the estimated t_{e} and c_{e}^2 for other throughput levels than the throughput level for which the arrival and departure data were collected. It is also valid to add new machines to the aggregate model, to study how the manufacturing system would respond to such an increase of capacity.

However, in many practical situations, the FIFO assumption does not apply. Even worse, machines may stay idle while processing is carried out at other machines (violation of the non-idling assumption). In these situations, the algorithm by Jacobs, Etman, van Campen, and Rooda (2003) can still be used to account for these dispatching related time losses in the EPT (this was actually the original motivation for the paper by Jacobs, Etman, van Campen, and Rooda (2001)). A complication is that a lot that generates an EPT start at a certain machine is not necessarily processed at this particular machine. There are no unique

pairs of EPT start and EPT end time anymore. For certain departures the EPT start time due a different machine has to be used in the EPT calculation if one wants to account for *all* dispatching related capacity losses. The selection of the pairs of EPT start and EPT end times is sometimes a rather arbitrary choice. From a *mean* effective process time point of view the way of accounting makes no difference. However the choice of the start time does affect the estimated variability. Our experience is that the estimated c_e^2 value may become too large, leading to a too high predicted cycle time. Also, a certain degree of utilization dependency of the estimated parameters is introduced, see (Wu and Hui 2008).

Alternatively, the dispatching related losses are not included in the EPT, but modeled separately. That is, the dispatching rule is considered as a control rule which is excluded from the aggregation. We return to this issue in Section 6.

Jacobs, Etman, van Campen, and Rooda (2003) illustrated their EPT measurement approach for a range of single-lot-processing workstations at the (former) Philips Semiconductors' MOS4YOU wafer fab. Arrivals and departures were collected from the manufacturing execution system (MES). The computed mean effective process time t_e and the squared coefficient of variation c_e^2 are depicted in Figure 2 (pictures from: Jacobs, Etman, van Campen, and Rooda (2003)). The figure clearly shows that the mean effective process time is larger than the mean natural processing time, and that the effective variability present at the workstation is substantially larger than the nominal value c_0^2. Figure 2(c) presents the cycle time factor observed from the MES and predicted using (1). The cycle time factor is the quotient of cycle time and nominal process time. The accuracy of the cycle time predictions lies for most of the workstations within 15% of the cycle times observed in the fab.

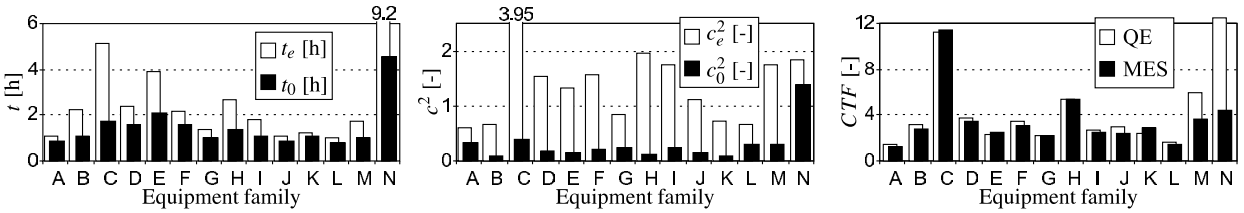


Figure 2: EPT-based *G*/*G*/*m* models of single-lot workstations at the former Philips Semiconductors MOS4YOU wafer fab: (a) measured t_e (b) measured c_e^2 (c) cycle time factor (measured and predicted).

4 WIP-DEPENDENT EPTs

Many equipment in semiconductor manufacturing carry out a series of process steps at various chambers inside the tool. There is a flow of wafers inside the machine. Wafers of multiple lots may be in process at the same time. For such integrated processing tools, the *G*/*G*/*m* approximation (1) may become inaccurate; the approximation does not account for the multi-processing behavior. Wafers from a new lot may already start processing before the wafers of the previous lot have finished.

Our efforts to better account for the multi-processing behavior, resulted in two fundamental changes in the aggregate modeling concept (Kock, Etman, Rooda, Adan, van Vuuren, and Wierman 2008): (a) make the EPT distribution in the aggregate model *workload dependent*, and (b) measure this workload dependency from the stochastic fluctuations of the workstation in operation. The idea is that the work-in-progress (wip) of any stochastic processing unit varies in time. If for every lot we register besides the effective processing time also the wip the lot experienced during its visit, we can sort the EPT realizations in accordance to the wip-level. This provides us with wip-dependent EPT distributions that we can use to build the workload dependent aggregate model. Note that this concept is different from the approach by Rose (2007) who built inventory-dependent aggregate delay distributions by running a detailed simulation model at various utilization levels.

We assume again a *G*/*G*/*m* type of aggregate model, with the difference that the mean and variance of the process time distribution depend on the number of lots in the system. Obviously, we do not have an analytical expression like (1) anymore, so we build an aggregate simulation model. We limit the number of wip-levels in the aggregate model to N, assuming that for N or more lots in the system, the system has become approximately load independent, indicating that the system is operating near its maximum throughput. In the aggregate model the process time for a certain wip-level is modeled using a distribution function that fits well to the measured EPT distribution. In our work we have mostly used Gamma distributions, mainly because its two parameters are determined by the mean and variance. But any other, possibly more suitable, distribution function can be used in the aggregate model, also distributions with more than two parameters.

For every wip level $0 \leq w \leq N$ we need an EPT-distribution. The method to determine EPT realizations from arrival and departure times is essentially the same as described in the previous section, apart from the assignment of EPT start times to *lots* instead of *machines*. We take the collected arrival and departure data set, and look at the data from the aggregate model point of view. Then, a new EPT starts in either of the following two cases (Veeger, Etman, van Herk, and Rooda 2010b):

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

Recall here that we take an aggregate model point of view, with servers that process one lot at a time, while the physical machine in the real workstation may process multiple lots at the same time. So upon EPT start the lot may already be physically in process. An EPT ends when:

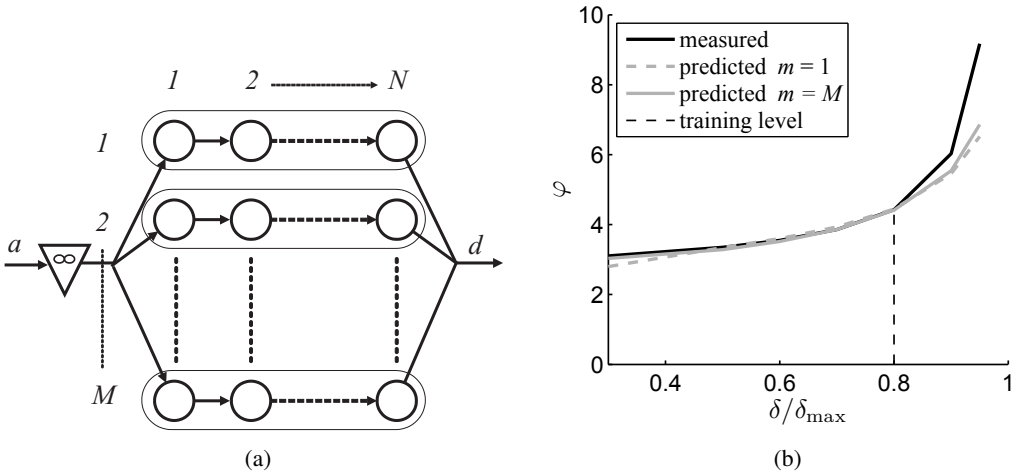
- a lot departs.

Upon the departure of lot i, the EPT of lot i is obtained by subtracting the EPT start time of lot i from the departure time. This EPT realization is assigned to the wip-level that corresponds to the number of lots present in the system upon the EPT *start* of lot i (including lot i).

There is one exception: a departing lot i may have overtaken m or more lots. In the aggregate model, upon its arrival no EPT start is assigned since m or more lots are already in the system. As a result, lot i has a departure time but no start time, and the EPT cannot be calculated. In that case Kock, Etman, Rooda, Adan, van Vuuren, and Wierman (2008) propose to use one of the EPT start times of other lots, for instance, the oldest one, the youngest one, or a random one. The EPT start time that is used is then reset and immediately re-started.

The aggregate model with wip-dependent EPTs appears to be remarkably accurate. What is more, instead of an m server aggregation of a workstation with m machines, also a *single* server aggregation of the m machine workstation can be considered. The *single* server aggregation appears to provide an approximation of almost equal accuracy in many cases. Consider for instance the fictitious multi-processing workstation depicted in Figure 3(a). At each workstation lots flow through a series of process steps (for ease of modeling we have assumed a flow of lots instead of a flow of wafers that belong to multiple lots). For this fictitious workstation, we have carried out an extensive simulation study to investigate the accuracy of the single server and the m server aggregation. We have investigated the influence of the number of integrated processes N, the number of parallel machines M, the workstation utilization level during the period of arrival/departure data collection, and the variability at the workstation, amongst others (results are yet to be published). Figure 3(b) depicts the case $N = 3$, $M = 4$, utilization $\delta/\delta_{\max} = 0.8$ (fraction of arrival rate and maximum throughput), Poisson arrivals, exponential process times, and a random pick rule (see previous paragraph). The figure shows that both the $m = 1$ and the $m = M$ aggregate model quite accurately predict the mean cycle time in a wide throughput range. For utilizations $\delta/\delta_{\max}$ substantially higher than

the utilization during the EPT data collection period (training level), the accuracy of the predicted mean cycle time decreases.



(a) (b)

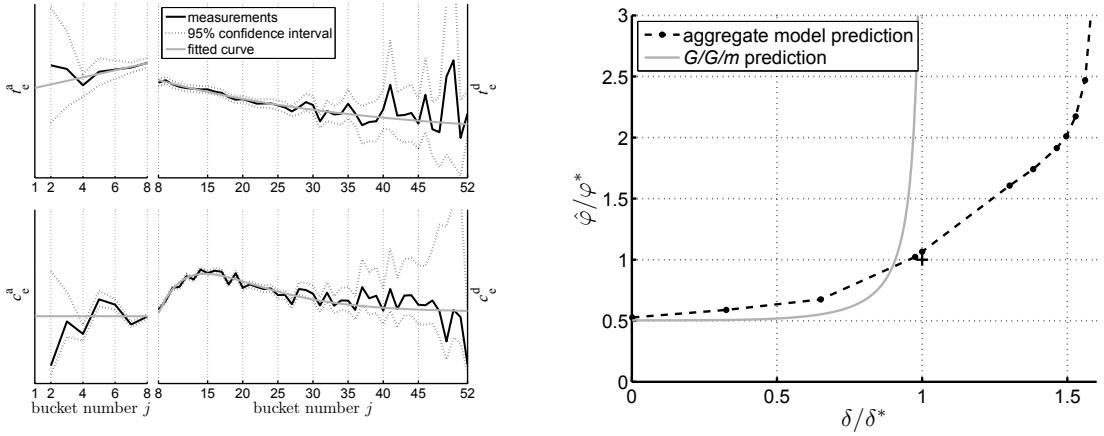
Figure 3: Aggregate model with wip-dependent EPTs applied to a fictitious multi-processing workstation: (a) workstation setup (b) cycle time-throughput curves.

In industry practice the number of arrival and departure events that can be collected in a certain time period is limited. As a consequence, for wip-levels that rarely occur during the data collection period only very few EPT-realizations will be obtained. Semiconductor equipment are typically quite heavily loaded, which means that for wip levels close to zero usually no observations are obtained. To overcome this difficulty, Veeger, Etman, van Herk, and Rooda (2010b) introduced a curve fitting procedure. They distinguish between EPTs started upon arrival, and EPTs started upon departure, and fit empirical closed form expressions to the t_e and c_e^2 as a function of the wip. An example from their paper is depicted in Figure 4(a). The horizontal axis denotes the wip level at the CD-measurement station. The left hand side of the picture refers to EPTs started upon arrival, and the right hand side to EPTs started upon departure. Clearly the behavior is different for these two cases. In Figure 4(b) the cycle time - throughput curve predicted by the *m*-server wip-dependent aggregate model is depicted and compared with the curve obtained from a 'classical' *G*/*G*/*m* approximation fed with wip-independent t_e and c_e^2 parameter values. The throughput on the horizontal axis is represented by the fraction of the arrival rate δ and the throughput δ^* at the training level. Similarly the cycle time on the vertical axis has been scaled for reasons of confidentiality. Thus, the cross depicts the cycle time observed at the training level (operating point). The wip-dependent effect at the workstation appears to be quite pronounced. The classical *G*/*G*/*m* approximation is inferior compared to the wip-dependent approximation.

5 WIP-DEPENDENT EPTs AND OVERTAKING

The third step we have made is to extend the wip-dependent aggregate model such that it can predict the cycle time *distribution* instead of only the *mean* cycle time. This was motivated by the desire to develop simplified models for the analysis of on-time delivery performance of equipment groups. For the calculation of cycle time distributions, detailed simulation models are almost exclusively used.

The new element introduced in the aggregate model is a wip-dependent overtake distribution which determines the *order* in which lots are processed (Veeger, Etman, Lefeber, Adan, van Herk, and Rooda 2011). We assumed a *single* server aggregation; that is, all overtaking observed in the modeled system, including overtaking due to processing at parallel machines, is accounted for in the overtake distribution



(a) (b)

Figure 4: Aggregate model with wip-dependent EPTs and curve fitting applied to an integrated processing workstation from Crolles 2 wafer fab in France: (a) measured EPT distribution parameters and curve fits (b) cycle time-throughput curve.

of the aggregate model. Then, the overtake behavior can be easily reconstructed from the registered lot arrival and departure times.

The aggregate model is visualized in Figure 5 (b). The queue in the new aggregate model is *not* a normal queue, but contains *all* lots that are currently in the system, irrespective of whether they are waiting or in process. Upon arrival of a new lot, the number of lots to overtake K is sampled from an overtake distribution. Also, the server is not a physical server, but a timer. The process times of this timer are sampled from an effective process time distribution. When the sampled process time has elapsed, the lot that is currently first in the queue leaves the system. This means that for a departing lot that has overtaken all lots in the system, in the aggregate model the timer started before this lot arrived. A new process time starts (i.e. the timer starts) if a new lot arrives in an empty system, or if a lot departs leaving the system non-empty. The effective process time distribution and the overtake distribution are both assumed to be wip-dependent. Upon sampling from the distributions the momentary workload is needed.

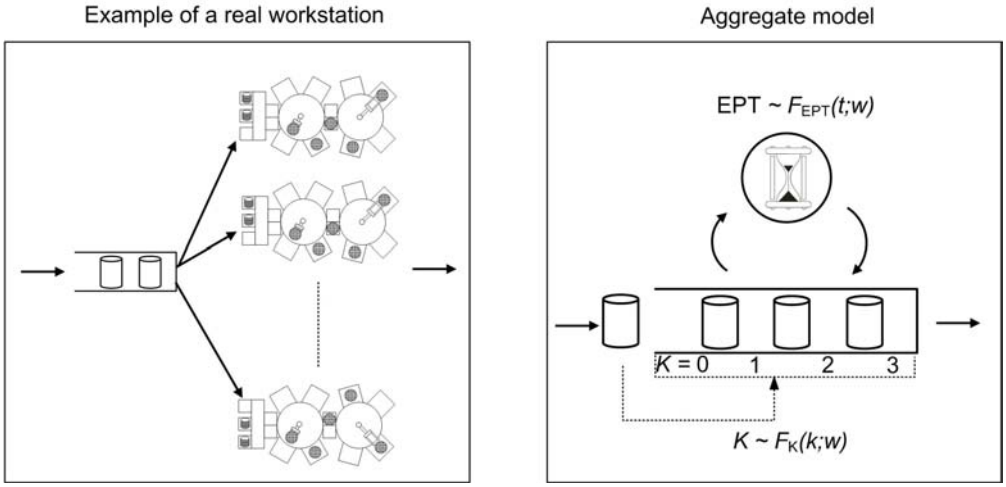


Figure 5: Aggregate model with wip-dependent EPTs and overtaking.

The wip-dependent EPT distribution and the wip-dependent overtake distribution are determined from arrival and departure data measured at the workstation in operation. We take again the aggregate model viewpoint. 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. Since the aggregate model consists of a single server (timer) only, every EPT start is subsequently followed by an EPT end. Thus, the EPT realization is obtained upon lot departure by simply subtracting the EPT start time from the departure time. Upon departure we also determine the momentary wip-level and the number of lots that were overtaken by the departing lot.

Veeger, Etman, Lefeber, Adan, van Herk, and Rooda (2011) illustrated the single server aggregate model with overtaking for a lithography workstation from the Crolles 2 wafer fab in France. EPTs were obtained for approximately forty thousand arrival and departure events extracted from the Manufacturing Execution System. Again curve fits were used to obtain closed-form expressions for the $t_e(w)$ and $c_e^2(w)$, see the left and middle plot in Figure 6. The measured overtake distribution (right hand plot in Figure 6) was used as is in the aggregate model. For wip levels below 18 and above 256 no EPT and overtake realization were obtained. In the empirical overtake distribution it was assumed that no overtaking takes place for wip$<$18; for wip$>$256, the same overtaking probabilities were used as measured for wip-level 256. Some predicted cycle time distributions are given in Figure 7. The right most picture corresponds to the training level, and shows that the prediction is reasonably accurate. Note that the tail of the distribution is quite long, which is disadvantageous for on-delivery performance. During the time frame of data collection, the utilization of the Lithography workstation was very high; therefore we were interested to quantify the effect of slightly lowering the utilization of the workstation on the shape of the cycle time distribution, see the left and middle picture of Figure 7.

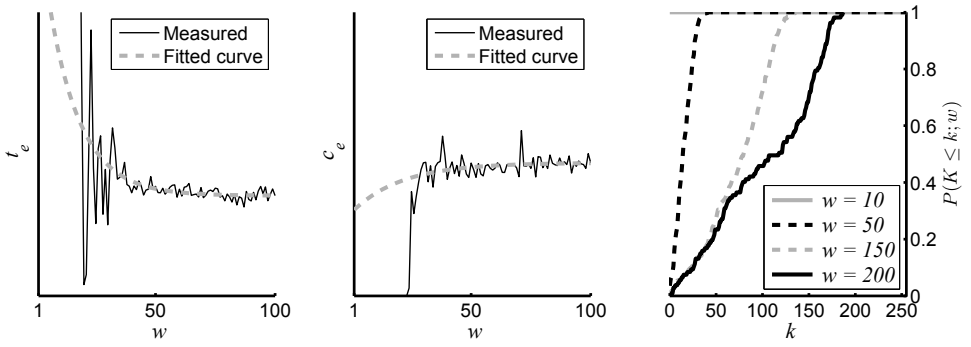


Figure 6: Aggregate model with wip-dependent EPTs, overtaking and curve fitting applied to the Lithography workstation from Crolles 2 wafer fab in France: measured EPT distribution parameters and curve fits (left, middle), and measured empirical overtaking distribution (right).

6 CONCLUSION AND OUTLOOK

We have developed a new modeling approach to build simple but accurate models of manufacturing systems by means of aggregation without the need to quantify the aggregated details. The aggregate model parameters are estimated directly from arrival and departures times. Figure 8 depicts our aggregate modeling framework. The square box at the top represents the modeled system; the square box at the bottom represents the aggregate model. The modeled manufacturing system may be a machine, a workstation or a network of workstations; the aggregate model is the desired abstraction from reality, with only the essential characteristics modeled explicitly and the other details aggregated in the effective process time distributions of the aggregate server(s). The parameters of the effective process time distributions are estimated from arrival and departures times measured from the manufacturing system in operation. The aggregate model

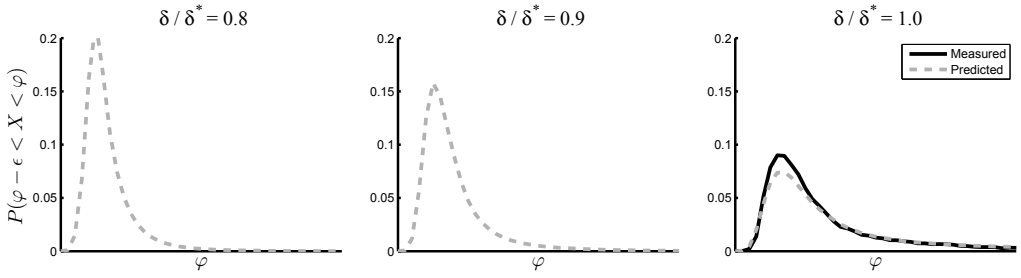


Figure 7: Aggegregate model with wip-dependent EPTs, overtaking and curve fitting applied to the Lithography workstation from Crolles 2 wafer fab in France: measured and predicted cycle time distributions for various throughput levels, with δ^* the throughput of the workstation in operation for which the EPT measurements have been obtained.

that is trained at a single utilization level is used to predict the cycle time performance at other utilization levels. No detailed modeling is used to build the aggregate model.

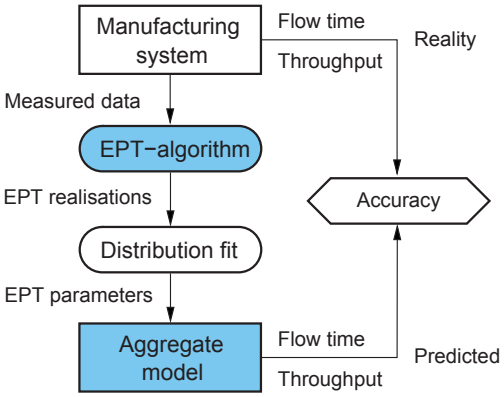


Figure 8: EPT-based aggregate modeling framework.

In our work, we have mainly focused on modeling workstations in the context of semiconductor wafer fabrication. The aggregate models we have used are: a closed-form *G*/*G*/*m* queueing expression, a *G*/*G*/*m* alike simulation model with wip-dependent process times, and a single server simulation model with wip-dependent inter-departure times and overtaking. The accompanying algorithms to estimate the aggregate model parameters from arrival and departure times have been developed. To this end we take the aggregate model view on the observed arrival and departure data. This means that the measured effective process time distribution should be interpreted as such.

We are not limited to these three types of aggregate models. Many more options are possible. Depending on the purpose of the model, one may want to explicitly incorporate specific characteristics in the aggregate model. If one wants to use the aggregate model for optimizing tool configuration, then the aggregate model should explicitly include the optimization variables of interest. For instance Veeger, Etman, van Herk, and Rooda (2010a) wanted to predict what will happen to the cycle time - throughput operating curve if the mix changes. Therefore they included the recipe qualification of machines in their aggregate model. Another example is batching machines, see Jacobs, van Bakel, Etman, and Rooda (2006). Note that adding details may require additional factory data to be collected, also data other than just arrival and departure times. Obviously, adding details comes at a price and the advantages of aggregation may diminish.

Neither are we limited to modeling workstations. The aggregate model may be also a network of machines or workstations. For instance, Kock, Wullems, Etman, Adan, Nijsse, and Rooda (2008) and

Kock, Etman, and Rooda (2008) used a finitely buffered flow line of machines as aggregate model for application in a car manufacturing assembly line. Key issue here is that blocking is an elementary property of a finitely buffered flow line; delays due to blocking should therefore *not* be included in the EPT of the aggregate servers in the flow line. Similarly, we may want to model explicitly certain control or dispatching rules whose contribution then has to be excluded from the EPT. See Lefeber and Armbruster (2011) for a discussion on modeling and control of manufacturing systems.

Instead of adding selected details, we may also increase the abstraction level of the aggregate model. Recall the single server representation of a physical multi-server workstation. But we may even want to go one step further: we may aggregate an entire network into for instance a single server aggregate representation. There is no free lunch here either; if the parameters of the aggregate model are utilization dependent, the more we aggregate the smaller the prediction range of the aggregate model becomes, since our model is trained at a single utilization level only. We have presented preliminary results of aggregating re-entrant flow lines at last year's Winter Simulation conference (Veeger, Etman, Adan, and Rooda 2010).

Another issue is the type of aggregate model. In Jacobs, Etman, van Campen, and Rooda (2003) we have used Sakasegawa's and Whitt's closed form expression; for the aforementioned finitely buffered flow lines we have developed efficient analytical models that can be fed with the measured t_e and c_e^2 values for the various machines in the line (van Vuuren and Adan 2009). But in all other cases we have relied on *simulation* to represent the aggregate model. It would be advantageous if also efficient analytical models with the desired characteristics such as wip-dependent behavior could be developed.

ACKNOWLEDGMENTS

This research was supported by the Technology Foundation STW, Applied Science Division of NWO and the Technology Program of the Dutch Ministry of Economic Affairs, and by NXP Semiconductors.

REFERENCES

Brooks, R. J., and A. M. Tobias. 2000. "Simplification in the Simulation of Manufacturing Systems". *International Journal of Production Research* 38 (5): 1009–1027.

Chandy, K. M., U. Herzog, and L. Woo. 1975. "Parametric Analysis of Queueing networks". *IBM Journal of Research and Development* 19 (1): 36–42.

Hopp, W. J., and M. L. Spearman. 1995. *Factory Physics: Foundations of Manufacturing Management*. 1st ed. New York: IRWIN/McGraw-Hill.

Hopp, W. J., and M. L. Spearman. 2008. *Factory Physics: Foundations of Manufacturing Management*. 3rd ed. New York: IRWIN/McGraw-Hill.

Jacobs, J. H., L. F. P. Etman, E. J. J. van Campen, and J. E. Rooda. 2001. "Quantifying Operational Time Variability: The Missing Parameter for Cycle Time Reduction". In *Proceedings of the 2001 Advanced Semiconductor Manufacturing Conference*, 1–10. Munich, Germany: SEMI/IEEE. doi: 10.1109/ASMC.2001.925605.

Jacobs, J. H., L. F. P. Etman, E. J. J. van Campen, and J. E. Rooda. 2003. "Characterization of Operational Time Variability using Effective Process Times". *IEEE Transactions on Semiconductor Manufacturing* 16 (3): 511–520.

Jacobs, J. H., P. P. van Bakel, L. F. P. Etman, and J. E. Rooda. 2006. "Quantifying Variability of Batching Equipment Using Effective Process Times". *IEEE Transactions on Semiconductor Manufacturing* 19 (2): 269–275.

Johnson, R. T., J. W. Fowler, and G. T. Mackulak. 2005, December. "A Discrete Event Simulation Model Simplification Technique". In *Proceedings of the 2005 Winter Simulation Conference*, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 2172–2176. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kock, A. A. A., L. F. P. Etman, and J. E. Rooda. 2008. "Effective Process Time for Multi-server Flowlines with Finite Buffers". *IIE Transactions* 40:177–186.

Kock, A. A. A., L. F. P. Etman, J. E. Rooda, I. J. B. F. Adan, M. van Vuuren, and A. Wierman. 2008. "Aggregate Modeling of Multi-Processing Workstations". Technical Report 2008-032, Eurandom, Eindhoven University of Technology, Eindhoven, The Netherlands. Accessed Oct. 16, 2011. http://www.eurandom.tue.nl/reports.

Kock, A. A. A., F. J. J. Wullems, L. F. P. Etman, I. J. B. F. Adan, F. Nijsse, and J. E. Rooda. 2008. "Performance Measurement and Lumped Parameter Modeling of Single Server Flow Lines Subject to Blocking: an Effective Process Time Approach". *Computers and Industrial Engineering* 54:866–878.

Lefeber, E., and D. Armbruster. 2011. "Aggregate modeling of manufacturing systems". In *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook, Volume 1*, edited by K. G. Kempf, P. Keskinocak, and R. Uzsoy, Volume 151 of *Springer International Series in Operations Research and Management Science*, Chapter 17, 509–536. New York, NY, USA: Springer-Verlag.

Morrison, J. R., and D. P. Martin. 2007a. "Performance evaluation of photolithography cluster tools". *OR Spectrum* 33:375–389.

Morrison, J. R., and D. P. Martin. 2007b. "Practical extensions to cycle time approximations for the *G*/*G*/*m*-queue with applications". *IEEE Transactions on Automation Science and Engineering* 4 (4): 523–532.

E. L. Norton 1926. "Design of finite networks for uniform frequency characteristic". http://www.ece.rice.edu/~dhj/norton/. Accessed Oct. 16, 2011.

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

Rose, O. 2007, December. "Improved Simple Simulation Models for Semiconductor Wafer Factories". In *Proceedings of the 2007 Winter Simulation Conference*, edited by S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 1708–1712. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sakasegawa, H. 1977. "An approximation formula $l_q = \alpha\beta^{\rho}(1-\rho)$". *Annals of the Institute for Statistical Mathematics* 29 (1): 67–75.

van Vuuren, M., and I. Adan. 2009. "Performance analysis of tandem queues with small buffers". *IIE Transactions* 41 (11): 882–892.

Veeger, C. P. L., L. F. P. Etman, I. J. B. F. Adan, and J. E. Rooda. 2010, December. "Single-server aggregation of a re-entrant flow line". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 2541–2552. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Veeger, C. P. L., L. F. P. Etman, E. Lefeber, I. J. B. F. Adan, J. van Herk, and J. E. Rooda. 2011. "Predicting Cycle Time Distributions for Integrated Processing Workstations: An Aggregate Modeling Approach". *IEEE Transactions on Semiconductor Manufacturing* 24 (2): 223–236.

Veeger, C. P. L., L. F. P. Etman, J. van Herk, and J. E. Rooda. 2010a. "Generating CT-TH-PM surfaces using EPT-based aggregate modeling". *Journal of Simulation* 4:242–254.

Veeger, C. P. L., L. F. P. Etman, J. van Herk, and J. E. Rooda. 2010b. "Generating Cycle Time-Throughput Curves using Effective Process Time based Aggregate Modeling". *IEEE Transactions on Semiconductor Manufacturing* 23 (4): 517–526.

Whitt, W. 1993. "Approximating the *GI*/*G*/*m* queue". *Production and Operations Management* 2 (2): 114–161.

Wu, K., and K. Hui. 2008. "The Determination and Indertermination of Service Times in Manufacturing Systems". *IEEE Transactions on Semiconductor Manufacturing* 21 (1): 72–82.

AUTHOR BIOGRAPHIES

L.F.P. ETMAN is an Associate Professor with the Manufacturing Networks group (formerly Systems Engineering group), Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. His current research interests include simulation-based optimization, structural and multi-disciplinary design optimization, and the effective process time method for aggregate modeling of manufacturing systems. His email address is l.f.p.etman@tue.nl

C.P.L. VEEGER was a PhD student with the Systems Engineering Group, Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. His PhD thesis work was on the development of the effective process time aggregate modeling method for semiconductor manufacturing applications. He is currently a consultant with OM Partners, Wommelgem, Belgium. His email address is CVeeger@ompartners.com

E. LEFEBER is an Assistant Professor with the Manufacturing Networks group (formerly Systems Engineering group), Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. His current research interests include modeling and control of manufacturing systems. His email address is a.a.j.lefeber@tue.nl

I.J.B.F. ADAN is Full Professor of the Manufacturing Networks group, Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. His current research interests include the analysis of multi-dimensional Markov processes and queueing models, and the performance evaluation of communication, production, and warehousing system. His email address is i.j.b.f.adan@tue.nl

J.E. ROODA is Professor Emeritus of the former Systems Engineering Group, Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. His research interests include design and analysis of manufacturing systems, manufacturing control, and supervisory machine control. His email address is j.e.rooda@tue.nl