Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

PREDICTING CYCLE TIME DISTRIBUTIONS WITH AGGREGATE MODELLING OF WORK AREAS IN A REAL-WORLD WAFER FAB

Patrick C. Deenen

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

Jelle Adan

Industrial Technology and Engineering Centre
Nexperia
Jonkerbosplein 52
Nijmegen, 6534AB, THE NETHERLANDS

John W. Fowler

Department of Supply Chain Management
Arizona State University
P.O. Box 874706
Tempe, AZ 85287-4706 USA

ABSTRACT

In a semiconductor wafer fabrication facility (wafer fab) it is important to accurately predict wafer outs, i.e. the remaining cycle time of the wafers in process. A wafer fab consists of multiple work areas, each containing a specific process technology, for example, photolithography, metal deposition or etching. Therefore, to accurately predict the wafer outs, an accurate prediction of the cycle time distribution at each work area is essential. This paper proposes an aggregate model to simulate each of these work areas. The aggregate model is a single server with an aggregate process time distribution and an overtaking distribution. Both distributions are WIP-dependent, but an additional layer-type dependency is introduced for the overtaking distribution. Application on a real-world wafer fabrication facility of a semiconductor manufacturer is presented for the work areas of photolithography, oxidation and dry etch. These experiments show that the aggregate model can, under certain circumstances, accurately predict the cycle time distributions in work areas by layer-type.

1 INTRODUCTION

This work is motivated by Nexperia's wafer fab in Manchester, United Kingdom. The wafer fabrication process starts with thin disks made of semiconducting material. The moving entities in a wafer fab are the lots, which in turn consist of a number of wafers. Each wafer contains thousands of integrated circuits (ICs) that are built layer by layer. From a functional point of view, work areas are the main building blocks of a wafer fab. An overview of the typical work areas in a fab is given in Figure 1. Each work area consists of several workstations. In turn, each workstation consists of one or multiple machines that provide similar processing capabilities. As the layers are built, a lot visits each work area numerous times, the precise number depends on the specific product.

As an illustrative example, a schematic overview of the oxidation work area is given in Figure 2. At Nexperia, this particular area consists of three different workstations, each of which constitute multiple parallel machines that share a single queue. From a scheduling point of view, this resembles a flexible flow shop. The wafer fab as a whole is modeled as a complex job shop. This is a job shop that has additional

978-1-6654-3311-2/21/$31.00 ©2021 IEEE

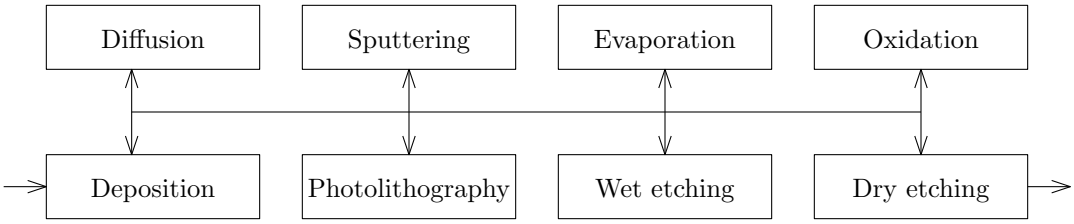


Figure 1: Main work areas in a wafer fab.

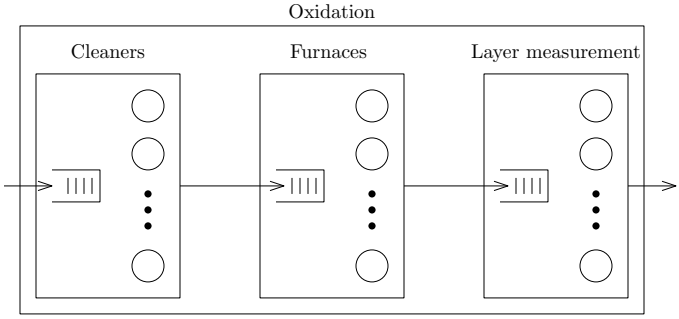


Figure 2: The oxidation work area as an illustrative example.

process restrictions such as sequence-dependent setup times, batching and re-entrant flows (Mason et al. 2002).

With such a highly complex manufacturing environment, production planning and control is especially challenging. From the customer's perspective, it is important to accurately predict when wafer lots will complete processing. The basis for this is an accurate prediction of the cycle time at each work area. For this purpose, simulation models with various levels of details are almost exclusively used (Mönch et al. 2012). Detailed simulation models enable accurate predictions but are computationally expensive. Model abstraction is necessary to reduce runtime and allow efficient experimentation (Dangelmaier et al. 2007). A common approach to abstract a detailed simulation model is aggregation, e.g. through modeling entire work areas by a constant delay or a simple single-server system. Usually, this is applied to non-bottleneck work areas while the bottleneck work area is still modeled in full detail (Rose 2007).

(Huber, Fowler, and Armbruster 2014) propose two algorithms to analytically approximate work in process WIP-dependent inter-departure times for tandem queues composed of a series of M/M/1 systems. (Kock et al. 2008) proposed an aggregate $G/G/m$-like queuing model with a work in process (WIP)-dependent aggregate process time distribution to predict the mean cycle time of (integrated) workstations. Here, WIP refers to the total number of lots at the workstation including the input buffer. The aggregate process time is referred to as the effective process time (EPT), a term first coined by (Hopp and Spearman 2011) and defined as "the time seen by a lot from a logistical point of view". The WIP-dependent EPT distribution can be calculated directly from actual arrival and departure data.

(Veeger et al. 2010a) demonstrated that the method of (Kock et al. 2008) suffices to predict the mean cycle time, but does does not accurately predict cycle time distributions. This limitation is largely a result of the First Come First Served (FCFS) rule in the aggregate model. A typical wafer fab processes a large mix of different products, on multiple machines of different ages and with various technologies. Consequently, machine speeds differ and tool dedication is often present, where dedication means that certain products cannot be processed on specific machines. Additionally, sequence-dependent setup times occur in several areas and some products may be prioritized before others. Hence, some lots are processed faster while others are delayed and the predicted cycle time distribution of the aggregate model with the FCFS rule is often too narrow. Evidently, a simple FCFS rule does not suffice to model this variability. Therefore,

(Veeger et al. 2010b) extended the aggregate model by taking into account the order in which lots are processed. Each lot that arrives in the aggregate model generally *overtakes* a number of lots already in the system. The number of lots it overtakes is sampled from a WIP-dependent overtaking distribution that can also be distilled from actual production data.

(Wu 2014) showed that the EPT can be a misleading concept and that it can introduce systematic errors. One of the conclusions was that the EPT is an explicit function of utilization and henceforth a dynamic quantity. This problem seems to be solved by introducing the WIP-dependent EPT distribution used by both (Kock et al. 2008) and (Veeger et al. 2010a). Another systematic error that can occur is due to time-based interruptions in the system, e.g. power outages or preventive maintenance. In this work (and even more elaborately in future work) it will be shown that this modelling approach is robust to moderate changes in operating conditions. As long as certain disturbances, such as short-term machine failures or maintenance activities, are also present in the time period of data collection, small and moderate disturbances will be accounted for in the model. Only if conditions radically change, i.e. long-term changes in capacity of highly utilized equipment, does one need to collect new data to fit the WIP-dependent EPT distributions. In some cases, one can simply add the length of the disturbance to the cycle time predictions for each lot.

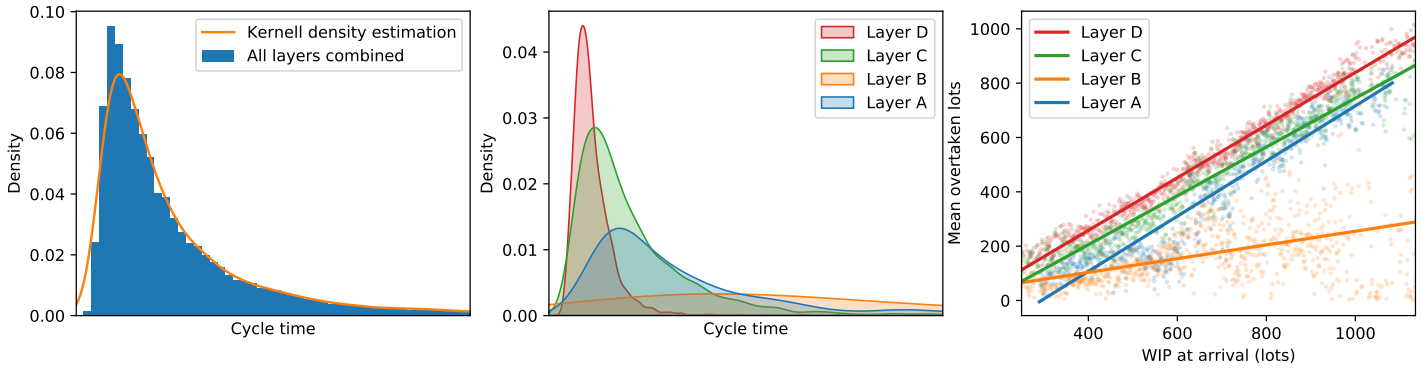
As mentioned, while products are built layer by layer, a lot visits each work area numerous times. The precise process sequence and requirements are product specific. However, though not identical, the layers of different product types are often quite similar in terms of process requirements. Hence, based on these similarities several different *layer types* are defined.

In Figure 3, the leftmost charts show a histogram and the corresponding kernel density estimation of the cycle times at the (a) photolithography, (b) oxidation and (c) dry etch work areas for the fab. The charts in the middle of Figure 3 show the kernel density estimation for the same data but segregated based on layer type. Through this segregation, it becomes clear that each layer type follows a different cycle time distribution. This is also reflected in the rightmost charts, which show the mean number of overtaken lots versus the WIP level with a first order fit, again segregated based on layer type. To accurately predict these different cycle time distributions amongst different layers, the aggregate model proposed in (Veeger et al. 2010b) does not suffice. Therefore, the aggregate model in this work is extended with an overtaking distribution which is not only WIP-dependent, but also layer-type-dependent.

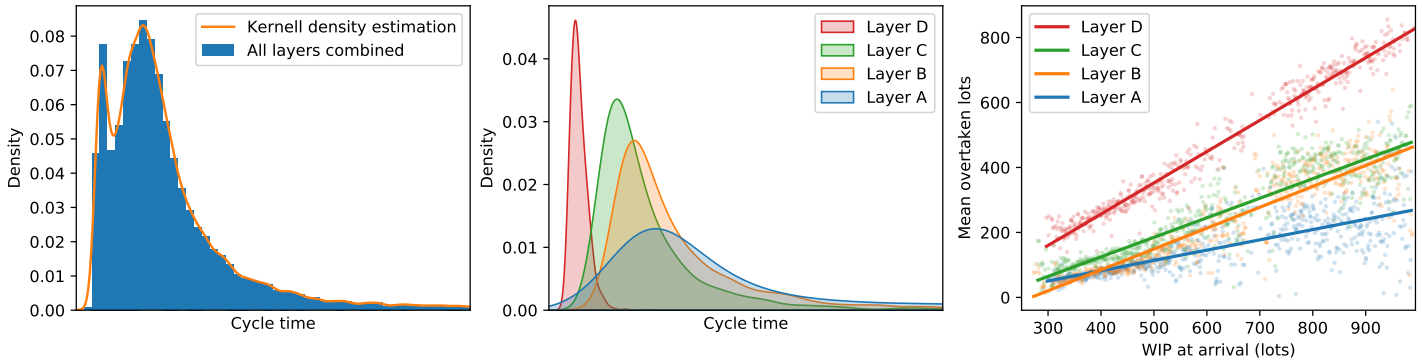
As it is our intention to model an entire wafer fab, the purpose of this work is to provide the building blocks for that. To realise these building blocks, two main contributions to the existing literature are made. The first one is that complete work areas are aggregated into one model. In contrast to (Veeger et al. 2010b), where single workstations were aggregated, a complete work area contains multiple sequential and parallel work stations. The second contribution is that a WIP-dependent Effective Processing Time (EPT) distribution similar to (Veeger et al. 2010b) is used, but additionally a WIP-dependent and layer-type-dependent overtaking distribution is introduced. The layer dependency in overtaking enables the aggregate model to accurately predict the different cycle times amongst different layer types. This is crucial when using the aggregate model as a block in a network, since the layer type mainly determines which work area it has to visit next.

2 MODELING APPROACH

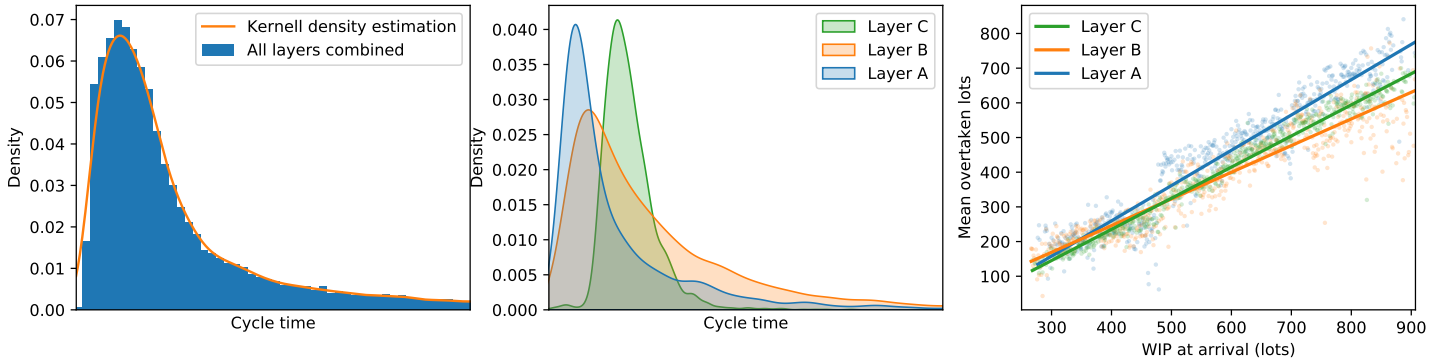
In this approach, an entire work area is modeled in an aggregate way. This aggregate model consists of a single server with an infinitely large buffer, a WIP-dependent processing time distribution, and an overtaking distribution that depends on both the WIP and layer type. Both distributions are determined from arrival and departure events measured at the work area. In this section, the aggregate model is briefly introduced and it is explained how the model parameters are determined.



(a) Photolithography.



(b) Oxidation.



(c) Dry etch.

Figure 3: Measured cycle time distribution for all layers combined (left), measured cycle time distribution per layer (middle) and mean overtaken lots per layer (right). The difference in cycle times amongst different layers can be clearly seen in the measured data. This motivates the introduction of a layer-type dependency in the aggregate model.

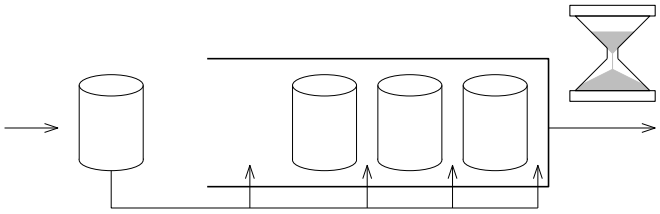


Figure 4: Schematic of the aggregate model. The number of jobs that an arriving job overtakes is sampled from a layer-type and WIP-dependent overtaking distribution. The processing time of a job, referred to as the EPT, is sampled from the WIP-dependent EPT distribution.

2.1 Aggregate Model

A schematic representation of the aggregate model initially proposed by (Veeger et al. 2010b) is given in Figure 4. In the current research, the aggregate model is used to approximate an entire work area (instead of a single workstation). This is done as follows: lots arrive in the queue according to some arrival process. This queue is not a queue as is common in queuing models, but contains all the lots that are in the system including the lots that are supposedly in service. When a new lot arrives, it is determined how many of the lots that are already in queue will be overtaken. This number is sampled from a so-called overtaking distribution. This probability distribution is both dependent on the number of lots already in the system upon arrival, as well as on the layer type of the arriving lot. The arriving lot overtakes the sampled number of lots and is placed in the queue. The server does not represent a physical server, but instead it is a timer that decides when a next lot leaves the queue. This timer starts when either (i) a lot arrives at an empty system or (ii) a lot departs and leaves a non-empty system behind. From the moment the timer starts all lots in the system can be overtaken by newly arriving lots, until the timer is elapsed. The time is sampled from a WIP-dependent distribution, where the WIP refers to the number of lots in the system after the arrival or departure has taken place. This time is referred to as an EPT. When the timer elapses, the lot that is first in the queue leaves the system.

2.2 Model Parameters

The input of the aggregate model constitutes a WIP-dependent EPT distribution and an overtaking distribution that is WIP and layer-type dependent. Both distributions are determined using actual arrival and departure data. For more details of the algorithm used to derive the distribution parameters directly from this data the reader is referred to (Jacobs et al. 2003). As stated before, the aggregate model is used to model an entire work area, e.g. the oxidation area depicted in Figure 2. Hence, for each lot, the time of arrival at the work area as well as the time of departure from the work area are collected. During this time span, the lot propagates through the workstations of the work area, and may undergo multiple state changes, as is shown in Figure 5. Also observe that an EPT starts when (i) a lot arrives at an empty system or (ii) a lot departs while at least one lot remains in the system. An EPT ends when a lot departs from the system. The corresponding WIP level of an EPT refers to the number of lots in the system at the start. Note that as the system is rarely ever empty, the EPT is practically the same as the the inter-departure time. A lot has overtaken another lot when it arrived later but departs earlier than the other lot.

The EPT-realizations are grouped according to the WIP level. For each group the mean and coefficient of variation are calculated. These are used to define the EPT-distribution at a particular WIP-level. In this case a gamma distribution is chosen as it appeared to describe the empirical data well. Similarly, the overtaking realizations are grouped according to WIP level upon arrival and layer type. These realizations are used as an empirical distribution. Note that it is assumed that all samples are independent and identically distributed (i.i.d.) random variables within the same WIP level (and layer type, in the case of the overtaking distribution). Although the individual EPT realizations are likely not truly i.i.d., this is an engineering approach to build the WIP-dependent EPT distributions, which are then used to predict the cycle time

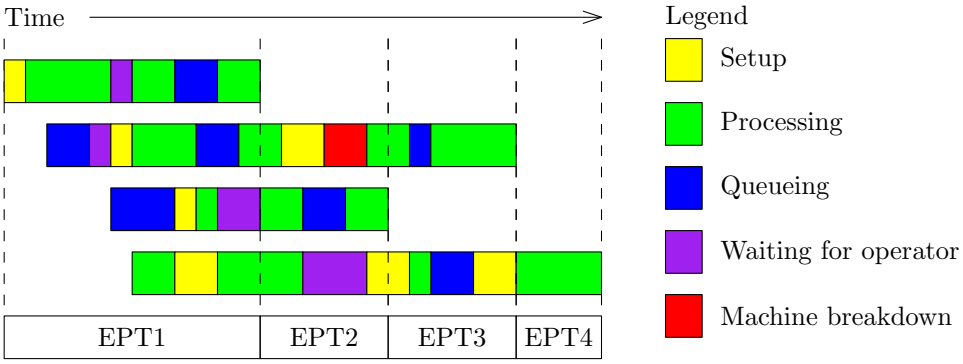


Figure 5: An illustrative example of the effective process time (EPT).

distributions. The results will show that these predicted cycle time distributions match well with the real data. In cases where the cycle time predictions do not match with the observed cycle times, this assumption might be causing the deviation.

There are two noteworthy issues with the use of actual production data. Firstly, the number of realizations for high and low WIP levels is usually very limited or not present at all. Secondly, the data is often quite noisy. For the WIP-dependent EPT distributions this is overcome by using a curve fitting procedure that approximates the measured mean EPT $t_e(w)$ for WIP level w by $\hat{t}_e(w)$. In particular, the following exponential function is used:

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

where θ represents the extreme at $w = \infty$, η represents the value of $\hat{t}_e(w)$ when $w = 1$ and λ is a so-called decay constant. Similarly, an exponential function of the same form as Equation (1) is used to approximate the coefficient of variation $c_e(w)$ as a function the WIP level w by $\hat{c}_e(w)$. The same issues are present in case of the overtaking distribution. However, there are no parameters that define the empirical distribution. Hence, to overcome the fact that there are no realizations for certain WIP levels, the realizations are divided into larger groups that cover a wider range of WIP levels. The separation based on layer type remains. For very high or low levels where there are no observations, samples are taken from the closest group in terms of WIP.

3 RESULTS

The proposed method is applied on three work areas in one of Nexperia's wafer fabs: photolithography, oxidation and dry etch. The data is obtained from the manufacturing execution (MES) system over a three-month period from August to October 2019. The data consists of timestamps recorded from all lots before and after processing on each machine, the so-called track-in and track-out activity. To retrieve the timestamps corresponding to the arrivals and departures at the work areas (as described in Section 2.1), some practical processing on the MES data is needed. For the sake of simplicity, this practical processing on the MES data is left outside the scope of this paper. In total, the arrivals and departures of 44168, 56437 and 33621 lots for, respectively, photolithography, oxidation and dry etch are collected. 87% of the lots contained the maximum of 25 wafers, the other 13% contained less than 25 wafers. The lots could be differentiated into 12 different layer types for photolithography and oxidation, and 11 different layer types for dry etch. All layer types are included in the aggregate simulation model. However, due to the large number of layer types, the figures in the analysis are limited to a selection of representative layer types.

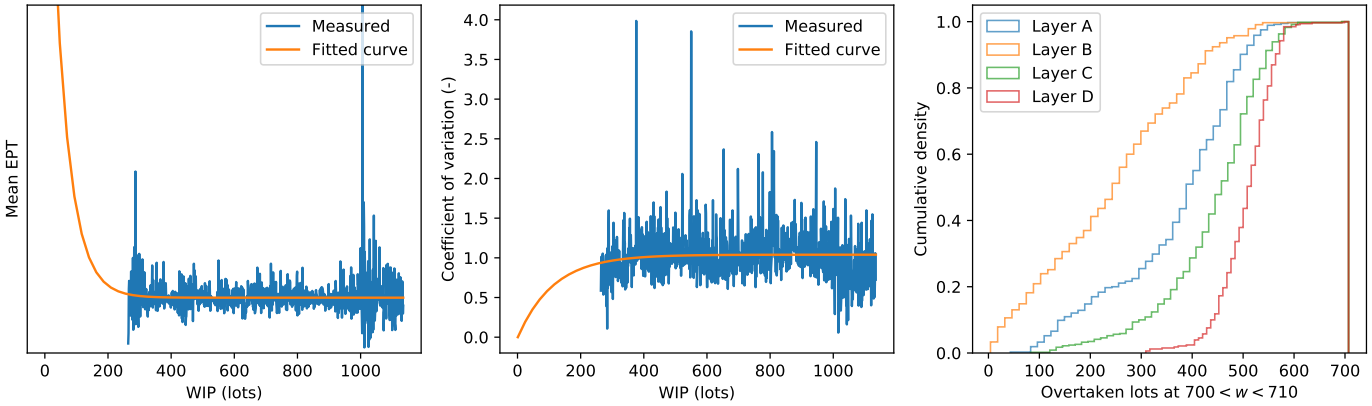
3.1 Model Parameters

The EPT and overtaking realizations are calculated from the arrivals and departures, as explained in Section 2.1. For the algorithms used for this, the reader is referred to (Veeger et al. 2010b). The observed mean and coefficient of variation of the EPT realizations of all layer types combined, t_e and c_e can be seen in the leftmost and middle plots of Figure 6. The rightmost plots show the cumulative overtaking distributions per layer type. A first observation one can make is that the coefficient of variation, c_e is significantly higher for the oxidation work area. As mentioned before, this area includes a step in which lots are batched and baked in a furnace. This batch processing along with relatively long processing times cause a high variability on the inter-departure of individual lots.

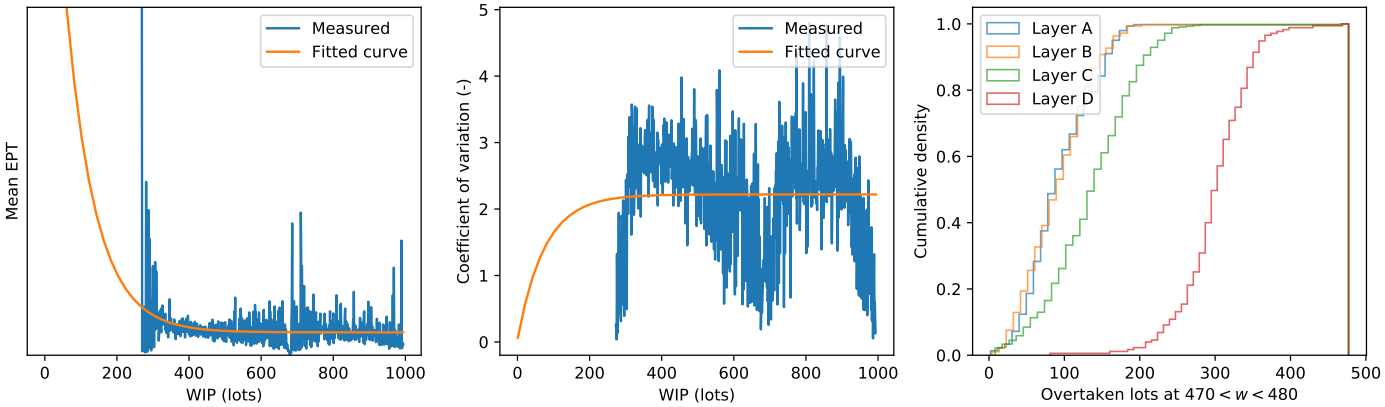
Another observation one can make, is that no data is available for low WIP levels. This is because complete work areas consisting of multiple workstations are aggregated. Even though certain tools might have run dry occasionally, complete work areas always had a WIP above 200. Another reason for the constant high WIP, is that these areas form the bottlenecks of this wafer fab and the fab has been running at high utilization to maximize the overall throughput. This also explains the fact that $t_e(w)$ does not decrease for the majority of the observed WIP range. This makes it challenging to identify the WIP-dependent behaviour. In particular, estimating η, λ and θ of Equation 1. Recall that η represents the value of $\hat{t}_e(w)$ at $w = 1$, θ represents the value of $\hat{t}_e(w)$ at $w = w_{max}$, with w_{max} the maximum WIP level observed, and λ represents the decay constant of the exponential curve. The same holds for the parameters of$\hat{c}_e$.

(Veeger et al. 2010b) proposed a nonlinear least-squares fitting procedure to estimate these variables. Due to the the aforementioned limited data in low WIP ranges, this procedure did not always yield parameters which could be used to accurately predict cycle times. To improve this, an educated guess is made to estimate η in $\hat{t}_e$ and in $\hat{c}_e$. The EPT realization of a lot in the case that $w = 1$, i.e. there is no other lot present, is equal to its cycle time in this area. The cycle time of such a lot does not contain any queuing and consists merely of the sum of all processing times and set up times on the machines and the transport between the machines. The estimate for the mean and variance of the processing and setup times is based on MES data containing track-in and track-out data and the transport times are estimated with the help of an experienced operator.

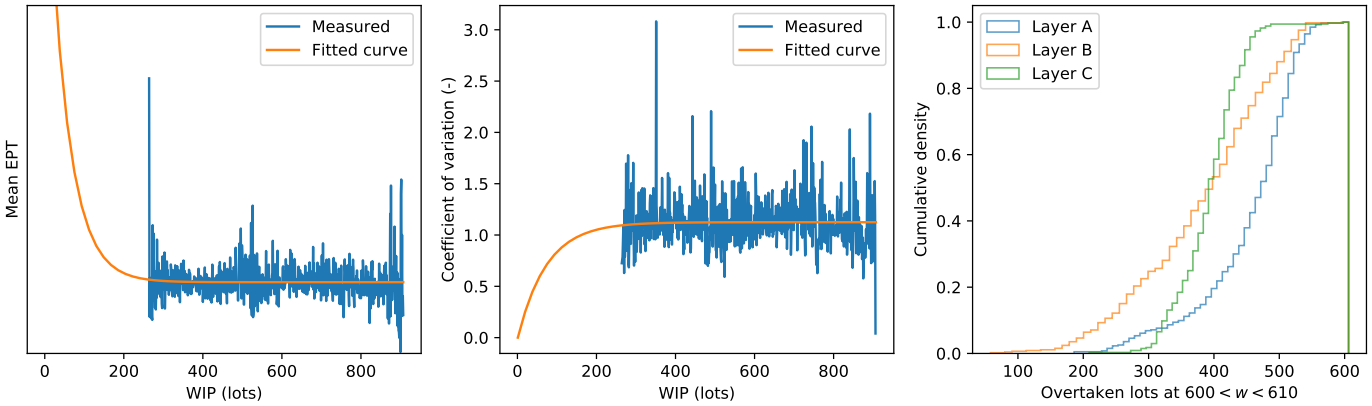
The cumulative overtaking probabilities for a certain WIP level per layer type can be seen in the right plots of Figure 6. Again, it is clear that certain layer types tend to overtake more lots than others, which motivates the introduction of layer-type-dependent overtaking. Since there is limited data and for some WIP levels there have been no observations, the data grouped in buckets with a WIP range of 10 to construct the empirical overtaking distributions.



(a) Photolithography.



(b) Oxidation.



(c) Dry etch.

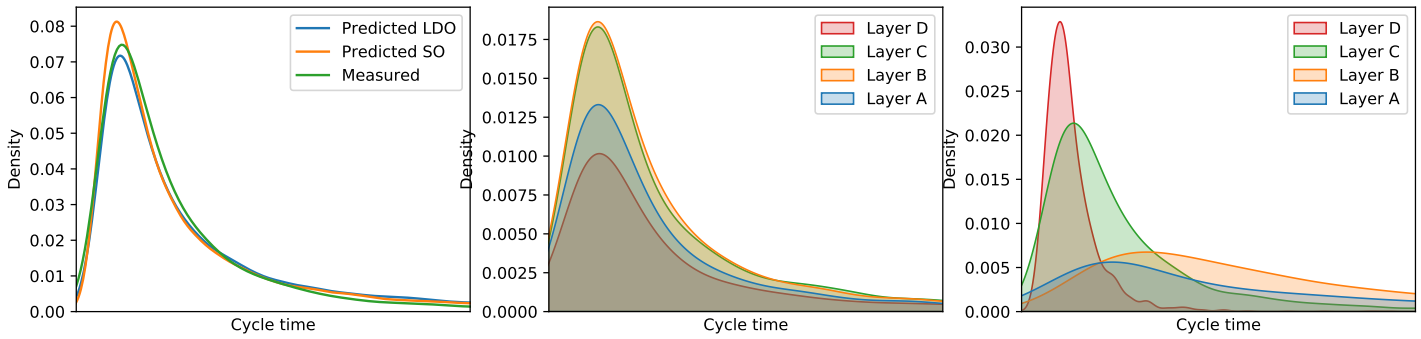
Figure 6: Measured and fitted mean EPT t_e (left) and coefficient of variability c_e (middle) for all layer types combined and cumulative overtaking probabilities per layer type (right).

3.2 Cycle Time Analysis

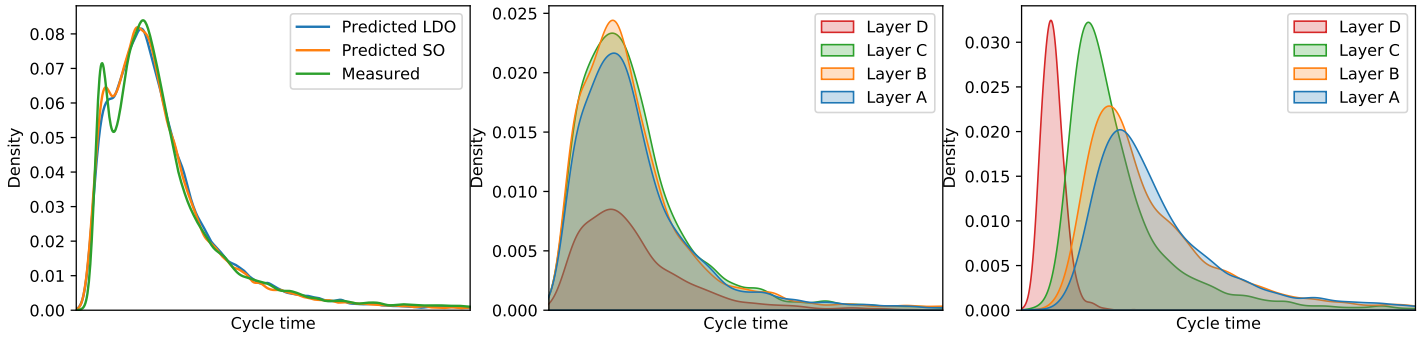
The aggregate models as depicted in Figure 4 based on the fitted values for $\hat{t}_e$ and $\hat{c}_e$ are shown in Figure 6 and are used to estimate the cycle time distributions. Two versions of this model are being used: (1) a model with a single overtaking (SO) distribution and (2) a model with a layer-type-dependent (LDO) overtaking distribution. The simulation runs for two months, after a warm-up of one month, using similar arrivals as obtained from the MES data of August - October 2019. Depending on the work area this corresponds to 1 to $1.5 * 10^4$ lots for warm-up and 2 to $3 * 10^4$ lots for the remainder of the simulation.

The resulting predictions of the cycle time distributions are depicted in Figure 7. Although the peaks of the predicted distributions in dry etch and oxidation are slightly lower than the measured ones, it is clear that the model can accurately estimate the cycle time distributions, especially in the photolithography work area. As expected, the SO and LDO models have no significant difference when observing the cycle times of all layers combined. However, the benefit of LDO becomes apparent in the distributions of individual layers. Compared to the measured cycle time distributions per layer of Figure 3, the model with LDO is able to accurately estimate these as well. Recall that this result is based on the manual fitting procedure explained in previous section. Since this manual fitting procedure might be improved in future work, these cycle time estimates will become even more accurate.

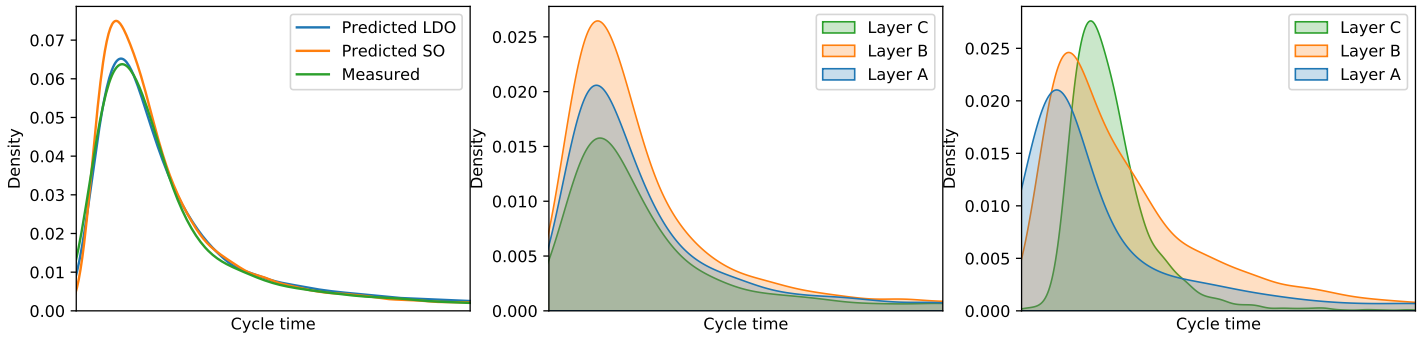
During the analyzed time period many disturbances, such as machine failures or maintenance activities, happened in the real-world fab. These results show that, since these disturbances were also reflected in the collected arrival and departure data, they are also accounted for in the model. Only if conditions radically change, i.e. long-term changes in capacity of highly utilized equipment, does one need to collect new data to fit the WIP-dependent EPT distributions.



(a) Photolithography.



(b) Oxidation.



(c) Dry etch.

Figure 7: Measured and predicted cycle time distribution for all layers combined (left), predicted cycle time distribution per layer with a single overtaking (SO) distribution (middle) and predicted cycle time distribution per layer with a layer-type-dependent overtaking (LDO) distribution (right).

4 CONCLUSIONS AND FUTURE WORK

This paper represents a first step for quickly estimating when the lots currently in the wafer fab will exit the system. In this step, a methodology has been developed that involves using historical arrival and departure events of lots in a given work area to develop an aggregate model of the work area. The aggregate model uses an empirical distribution of the effective processing time (EPT) in an area by WIP level and an empirical distribution of the amount of overtaking that takes place in the work area based on the WIP level when the lot enters the area and the layer that it is on. The experimentation in this work indicates that the approach is effective for the photolithography, oxidation, and dry etch areas of Nexperia's Manchester fab. However, it is noted that finding the correct values for t_e and c_e might be challenging, especially because data on low WIP levels are not available for complete work areas.

There are a number of areas that will be investigated going forward. This work showed that the proposed modelling approach gives an accurate cycle time prediction for the same data set as that the model is trained on. In future work, it will be investigated how well it performs on other data sets and under which circumstances the predictions will deviate. Furthermore, the estimation of the parameters which determines the behaviour for WIP ranges where no data is available, will be be refined and an automation for this process is sought. Finally, models of each of the fab areas will be constructed and connected together to estimate the time that lots will exit the fab, which is the ultimate goal. Once the methodology to model the entire fab is complete, the possibility will be investigated of using the same methodology with a simulation model to analyze different dispatching policies including using the estimates from the area models in look ahead and look behind estimates in the dispatching rules.

REFERENCES

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

Hopp, W. J., and M. L. Spearman. 2011. *Factory physics*. Waveland Press.

Huber, D., J. Fowler, and D. Armbruster. 2014. "Simplification of DES models of M/M/1 tandem queues by approximating WIP-dependent inter-departure times". *Simulation* 90(10):1188–1196.

Jacobs, J., L. Etman, E. Van Campen, and J. Rooda. 2003. "Characterization of operational time variability using effective process times". *IEEE Transactions on semiconductor manufacturing* 16(3):511–520.

Kock, A., L. Etman, J. Rooda, I. Adan, M. Van Vuuren, A. Wierman et al. 2008. "Aggregate modeling of multi-processing workstations". *Eurandom, Eindhoven, The Netherlands, Ext. Rep* 32.

Mason, S. J., J. W. Fowler, and W. Matthew Carlyle. 2002. "A modified shifting bottleneck heuristic for minimizing total weighted tardiness in complex job shops". *Journal of Scheduling* 5(3):247–262.

Mönch, L., J. W. Fowler, and S. J. Mason. 2012. *Production planning and control for semiconductor wafer fabrication facilities: modeling, analysis, and systems*, Volume 52. Springer Science & Business Media.

Rose, O. 2007. "Improved simple simulation models for semiconductor wafer factories". In *2007 Winter Simulation Conference*, 1708–1712. Institute of Electrical and Electronics Engineers (IEEE).

Veeger, C., L. Etman, E. Lefeber, I. Adan, J. Van Herk, and J. Rooda. 2010b. "Predicting cycle time distributions for integrated processing workstations: an aggregate modeling approach". *IEEE Transactions on Semiconductor Manufacturing* 24(2):223–236.

Veeger, C., L. Etman, J. Van Herk, and J. Rooda. 2010a. "Generating cycle time-throughput curves using effective process time based aggregate modeling". *IEEE Transactions on Semiconductor Manufacturing* 23(4):517–526.

Wu, K. 2014. "Classification of queueing models for a workstation with interruptions: a review". *International Journal of Production Research* 52(3):902–917.

AUTHOR BIOGRAPHIES

PATRICK C. DEENEN is a doctoral candidate in the Department of Industrial Engineering of the Eindhoven University of Technology and a Sr. Business Process Analyst at Nexperia. His current research interests are in the area of modeling, control and optimization of manufacturing systems. His email address is patrickdeenen@hotmail.com.

JELLE ADAN is a doctoral candidate in the Department of Industrial Engineering of the Eindhoven University of Technology and a Sr. Business Process Analyst at Nexperia, Equipment and Automation Technologies (E&A). His current research interests are supply chain, manufacturing and chemical process optimization, and data mining. His email address is jelle.adan@protonmail.com.

JOHN W. FOWLER is the Motorola Professor of Supply Chain Management and recently served as Chair of the Supply Chain Management department in the W.P. Carey School of Business at Arizona State University. His research interests include discrete event simulation, deterministic scheduling, multi-criteria decision making, and applied operations research with applications in semiconductor manufacturing and healthcare. He has published over 130 journal articles and over 100 conference papers. He was the Program Chair for the 2002 and 2008 *Industrial Engineering Research Conferences*, Program Chair for the 2008 *Winter Simulation Conference (WSC)*, and Program Co-Chair for the 2012 *INFORMS National Meeting*. He was the founding Editor-in-Chief of *IISE Transactions on Healthcare Systems Engineering* and currently serves as a Healthcare Operations Management Departmental Editor. He is also an Editor of the *Journal of Simulation* and Associate Editor of *IEEE Transactions on Semiconductor Manufacturing* and the *Journal of Scheduling*. He is a Fellow of the Institute of Industrial and Systems Engineers (IISE) and served as the IIE Vice President for Continuing Education, is a former INFORMS Vice President, and served on the WSC Board of Directors. His email address is john.fowler@asu.edu.