Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

ENHANCEMENT OF SIMULATION-BASED SEMICONDUCTOR MANUFACTURING FORECAST QUALITY THROUGH HYBRID TOOL DOWN TIME MODELING

Patrick Preuss
André Naumann

D-SIMLAB Technologies GmbH
Wiener Platz 6
01069 Dresden, GERMANY

Wolfgang Scholl

Infineon Technologies Dresden GmbH
Königsbrücker Strasse 180
01099 Dresden, GERMANY

Boon Ping Gan
Peter Lendermann

D-SIMLAB Technologies Pte Ltd
8 Jurong Town Hall Road
#30-04 JTC Summit
Singapore 609434, SINGAPORE

ABSTRACT

Material flow forecast based on Short-Term Simulation has been established as a decision support solution for fine-tuning of Preventive Maintenance (PM) timing at Infineon Dresden. To ensure stable forecast quality for effective PM decision making, the typical tool uptime behavior needs to be portrayed accurately. In this paper, we present a hybrid tool down modeling approach that selectively combines deterministic and random down time modeling based on historical tool uptime behavior. The method allowed to approximate the daily uptime of reality in simulation. A generic framework to model historical down behavior of any distribution type, described by the two parameters Mean Time to Failure (MTTF) and Mean Time to Repair (MTTR) is also discussed.

1 INTRODUCTION

At Infineon Dresden two types of simulation-based decision support have been established: long-term simulation, with forecast granularity and horizon of weeks, and short-term simulation with a forecast granularity and horizon of days. Both approaches are different in terms of modeling abstraction and the derived output data granularity and as such they have a difference set of applicable use cases. An important modeling element are equipment downs.

For long-term simulation it is essential to model all expected down events as the forecast horizon can be up to a year and the complete amount of down time in the model is required to represent the fab capacity correctly. Long-term simulation is mainly used to give answers to strategic planning questions for the whole fab as well as on the equipment group level. Although the correct day of a down may not be predicted correctly, the average expected down percentage can be abstracted based on typical historical down behavior of each equipment group.

Short-term simulation requires a different modeling approach of equipment down behavior. This is due to the fact, that the forecast horizon is only up to two weeks and the granularity of the key performance indicators (KPI) is required on a daily basis. The majority of equipment downs are unscheduled downs and counter-based scheduled downs, for which it is not known in advance on which day a down event will happen. As such, if all downs are modeled purely on a random basis the chances are low to forecast the exact day of each down event. Besides that, in short-term simulation the rare atypical downs (i.e. downs that last much longer than expected) need to be removed from the modeling, because they are not very likely to happen during the one week forecast but can change the average weekly uptime significantly. To address the discussed issues, this paper proposes a down modeling approach that categorizing the downs as follows.

- Model downs with planning data
- Model downs as random distribution
- Model downs as constant daily factor
- Not model downs at all

2 MODELING APPROACH

In the previous implementation of the short-term simulation equipment downs have been modeled using an exponential distribution function for mean time to failure (MTTF) and a constant value for mean time to repair (MTTR). Both, MTTF and MTTR, are based upon pre-aggregated equipment down time behavior on the mainframe level. Further information about modeling system randomness can be found in (Law 2007).

Current analysis described in this paper reveals that single, exceptionally long down events (atypical events) can have a major impact on the MTTR values. Furthermore it has been shown that MTTF and MTTR distributions vary between the equipment groups and cannot be modeled appropriately using only exponential distribution or constant values. Additionally there is the requirement to enable the integration of preventive maintenance (PM) schedules and remove those PMs from the random distribution, in order to minimize stochastic effects in the short-term simulation.

Hence in the revised down modeling approach it is necessary to identify down events by their down event type, as well as by down duration, and eventually exclude certain down events from further processing. Therefore, instead of pre-aggregated values, the equipment's historical trace events have to be used and a possibility to represent more complex time to failure and time to repair distribution is required.

First step is a detailed analysis of equipment downs. Therefore historical downs of all tools within a specified timeframe were evaluated with the focus on typical duration of the down and interval of the down, the time between the end of a down until the start of the next down. Furthermore investigation on the statistical data mean, median, average deviation and the shape of the histogram were conducted. Other aspects evaluated are the duration of the historical trace horizon and the possibility to group equipment which show the same down behavior.

2.1 Down Classification

According to Infineon's uptime definition guidelines, equipment downs have been classified into three main categories: PM, scheduled downs, and unscheduled downs, whereby PMs can be further grouped into counter-based and time-based downs as shown in Figure 1. A similar classification is described in (Achermann 2008). The classification of downs was required because with regard to their cause the downs have a significant difference in duration and interval, so the quality of the modeled downs is better if the downs are analyzed and incorporated separately. Even though PMs are usually considered as scheduled downs, they have different characteristics compared to other scheduled downs. Maintenance work usually takes several hours and their occurrence is defined by a maintenance plan or a consumption counter. In addition the execution of PMs might be tied to the availability of qualified service personnel or spare parts which can increase the variance of the duration observed. Smaller scheduled downs, for example tool tests, only take a few minutes and occur more often. It is even possible that those downs, which are triggered by automatic tests, show a nearly constant duration as well as a constant interval.

Unscheduled downs on the other hand show a high variance with regard to duration and interval. Those downs can take only few minutes or up to hours and even days and the interval between two downs can also change over time.

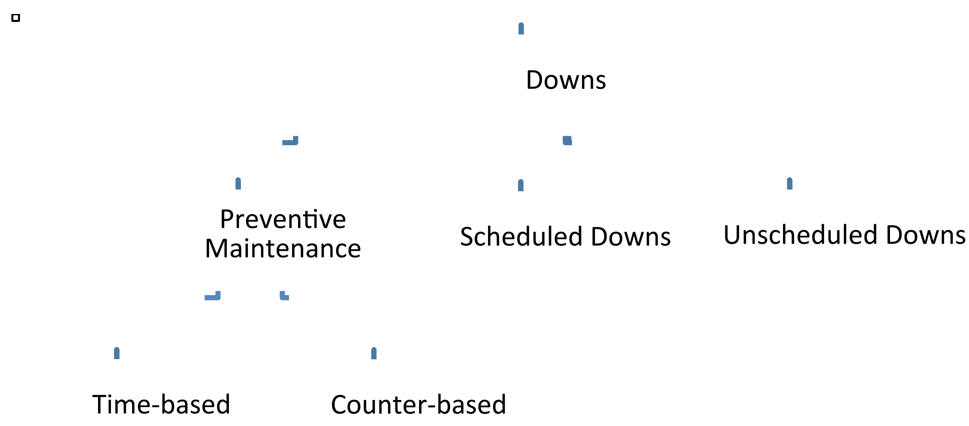


Figure 1: Categorization of equipment downs.

A second reason for the categorization was the targeted modular modeling concept which allows to manage each down type independently and replace randomly generated down data by deterministic modeling data whenever it is possible. Deterministic data can be either the creation of a new approach to model down information of a certain type or new planning data sources which become available. This concept allows a continuous model improvement, starting with stochastic modeling data and moving on to replace a subset of those data by deterministic information to further improve the forecast quality. Especially maintenance activities mostly follow a repeating interval and are executed within a specified time frame. If those information can be extracted or derived from data sources those maintenance activities do not need to be modeled using a stochastic approach.

Another scheduled down type are the counter based PMs. These maintenance activities are typically executed after a certain amount of wafers have been processed or a certain amount of equipment operating hours have passed. An approach on how counter based PMs can be incorporated into the model is described in Section 4.

2.2 Historical Down Event Processing

To process historical down event data, the time horizon for the data analysis had to be defined. The choice the of time horizon is a tradeoff between how up-to-date data are on the one hand and data completeness and significance on the other hand. If the time frame is too short, downs which happen rarely might not be tracked and the parameters derived do not represent the typical equipment down behavior. In addition the number of down event samples is small, which affects the statistical significance of the derived mean values and the reliability of the distribution. If the time frame is too long, changes in equipment's down time behavior will take a while until they are reflected in the down model.

So the starting point of the analysis was the maximum query time horizon of six months. To ensure that the chosen time horizon does not affect the quality of down behavior modelling, the down time behavior of the equipment in the past 6 months was investigated. As shown in Figure 2, the average uptime behavior of the equipment does not vary for the past 6 months when atypical events are excluded. Applying this same strategy to all equipment groups, we determined the time horizon to be considered was 6 months.

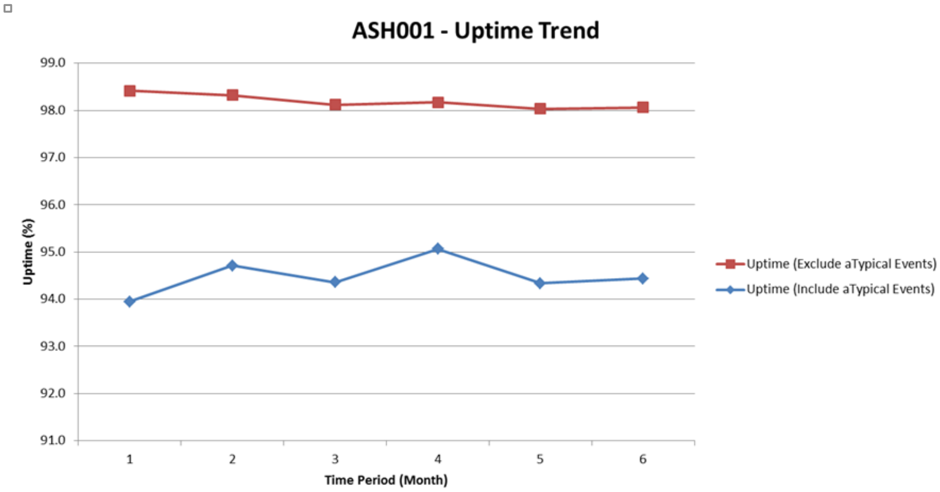


Figure 2: Uptime trend for an equipment using an analysis period between one and six months.

To further increase the amount of samples and therefore the statistical significance, equipment are merged into equipment groups where all equipment show the same down time behavior. Equipment of one manufacturer and conducting related processes are likely to have similar maintenance patterns and test intervals, similar wearing and utilization, and a similar probability of failure. The first digits of the equipment name served as the grouping criteria. It was analyzed whether all equipment within an equipment group are having same MTTR and MTTF, and whether the down duration and down interval distribution show the same characteristics as illustrated in Figure 3 below. All equipment groups had to undergo this analysis, whereby it was concluded that the equipment within a group indeed have a similar down time behavior.

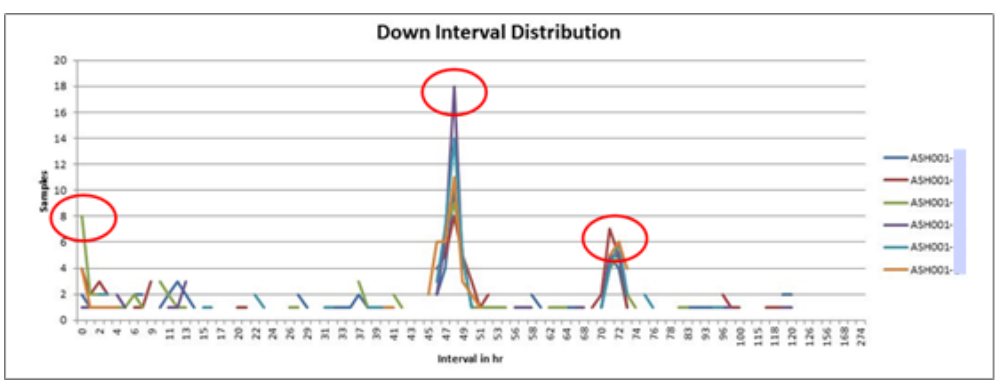


Figure 3: Down interval distribution of individual equipments within the group ASH001.

Due the way equipment downs are represented in the data sources, some data post-processing was required. An equipment down usually consists of several phases. For example, in the first phase an equipment is breaking down and waiting for repair, then the equipment is being repaired, followed by a phase of testing at the end. These steps are tracked by a sequence of multiple down states, but from modeling point of view they are only on single down, therefore a sequence of downs has to be joined, otherwise derived down durations and intervals would not represent the real down behavior. The down category of the combined down depends on the down which is the major cause. Combined downs containing unscheduled downs, will always be labeled as unscheduled downs. PMs followed by scheduled tests will be considered as PM as a whole.

A major part of the processing of historical down data is the removal of atypical down events. An atypical down is an excessively long down, which is not representative of the equipment's typical down behavior. These downs can have an huge impact on the equipment's MTTR. Even if the equipment has sufficient amount of samples, a single atypical down can multiply the MTTR, as illustrated in Table 1. Thereby the equipment would be modeled with a down time that is excessively long.

Table 1: A typical event example.

	Frequency	Duration (hr)	Average Duration (hr)
Down Stream A	50	1	1
Down Stream B	49	1	2.98
	1	100	

There are different criteria for the detection of atypical down events. On the one hand an absolute threshold can be defined, or on the other hand a relative value (multiple of mean) can be used. Whenever a down exceeds the threshold it has to be excluded from further usage.

2.3 Down Collections and Modeling Decision

After the processing of historical down events, down collections for each equipment groups are created. Each down collection contains down events of different down event types. For example one collection contains downs of all types (PM, scheduled and unscheduled down), while another collection contains only scheduled and unscheduled downs, but no PM. This separation is required to make sure that, for example, the latter down distribution will be used if PM plan is available for the equipment group.

Each down collection of an equipment group will be created, only one down collection will be used in a model. The redundancy of providing multiple potential down modeling data is to address varying availability and up-to-date state of data sources. With this method down modeling data can be swapped with each model without the need of rerunning the historical down event analysis.

After the down collections are created, we have to decide how the downs are going to be modeled for each collection. There are four different approaches: (i) not modeling the downs, (ii) using constant MTTR and MTTF, (iii) using distribution functions for MTTR and MTTF and (iv) using plan data. A fifth approach, dynamic down triggering, is described in the Section 4. Which approach is applicable for which down category is shown in Table 2.

Table 2: Modeling approaches based on down category.

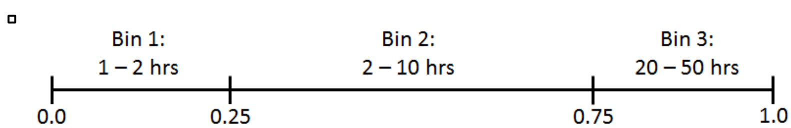
	PM		Scheduled	Unscheduled
	Counter-based	Time-based		
Not Modeling	x	x	x	x
Constant	x	x	x	x
Distribution	x	x	x	x
Plan		x	x	
Dynamic	x			

The first assessment to be conducted is to check whether downs happen frequent enough to be modeled. Insufficient samples will restrict the statistical significance of computing MTTF and MTTR values. Besides the number of samples, it is also important to assess the number of days during which downs will be observed within the forecast period. Analysis revealed that if less than 50% of the days are affected by a down, it is better not to model them. A more detailed insight into how this quality measurement has been done is described in Section 4.

If there are enough samples and sufficient forecast days are affected by down events, the down is modeled as either constant capacity reduction or through a statistical distribution. The deciding criteria is to assess the probability that the down will be generated on the right day in the simulation model. This assessment can be done by analyzing the spread of down intervals. A big spread leads to a high uncertainty and stochastic effects, subsequently a low probability to match the correct day with a random generated down. Therefore these downs will be modeled by a constant MTTF and MTTR to reduce the deviation between the uncertain real down events and the downs modeled in the simulation. If the spreading is low, then a distribution will be used to model these downs and the required parameters for the probability density functions for MTTF and MTTR will be determined.

2.4 Nested-Uniform-Distribution Custom Extension for AutoSched AP

The simulation engine used in this project is AutoSched AP. It provides a limited set of distribution functions. This triggers us to introduce a customized distribution function, called Nested-Uniform-Distribution, to model any kind and shape of distribution. To model a distribution, multiple bins are defined. Each bin contains a range of values and a probability. Using a uniformly distributed random number, a bin is selected and subsequently a value from within the bin's range is generated, as illustrated in Figure 4.



MTTF Range (hrs)	Probability
1 – 2	0.25
2 – 10	0.50
20 – 50	0.25

Figure 4: Bin definition for Nested-Uniform-Distribution.

Depending on the number of bins, the granularity of the modeled distribution can be adjusted in a flexible manner. As shown in Figure 5 the developed simulator custom extension is able to portray the historical down interval distribution curve.

3 RESULTS

In our study we chose several simulation models equally spread over an observation period of one month and compared the results for three different equipment groups ASH001 (reliable tool group with a high uptime), ETC010 (large fluctuation in the daily uptime) and MES075 (high impact of atypical events). The uptime comparison charts for each of the group are shown in Figures 6, 7 and 8.

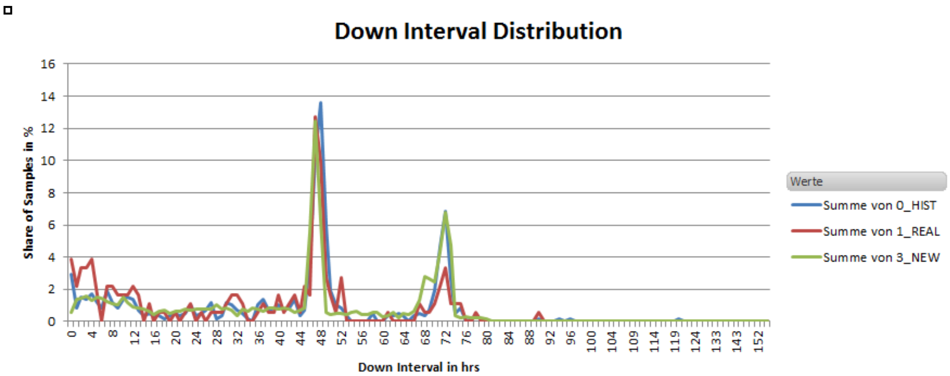


Figure 5: Down interval distribution curves of historical observation period (0), real data of simulation forecast period (1) and simulator output (3).

As seen in the figures, the new down time modeling approach is able to represent the real uptime much better than the previous approach. This improvement is contributed by the enhanced down event data processing and the generic down distribution custom extension as described in Section 2.

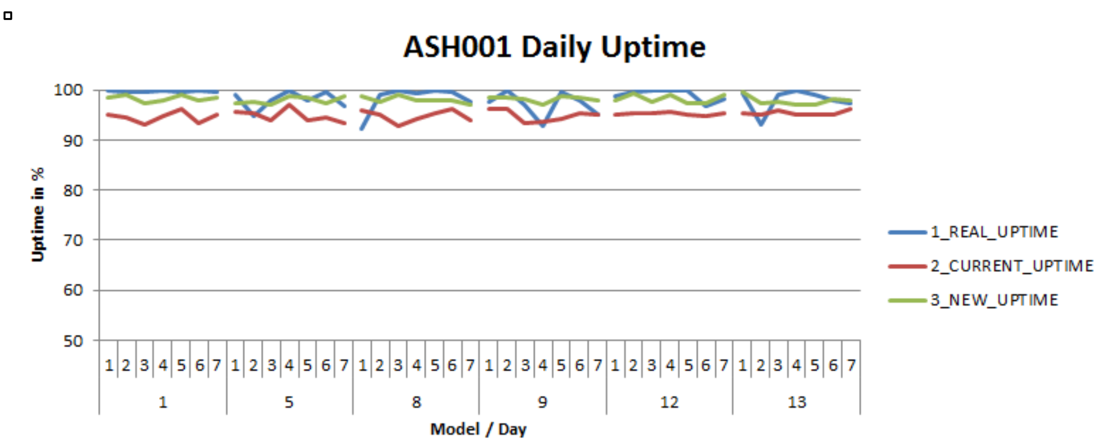


Figure 6: ASH001 daily uptime curves of real uptime (1), old approach (2), new approach (3).

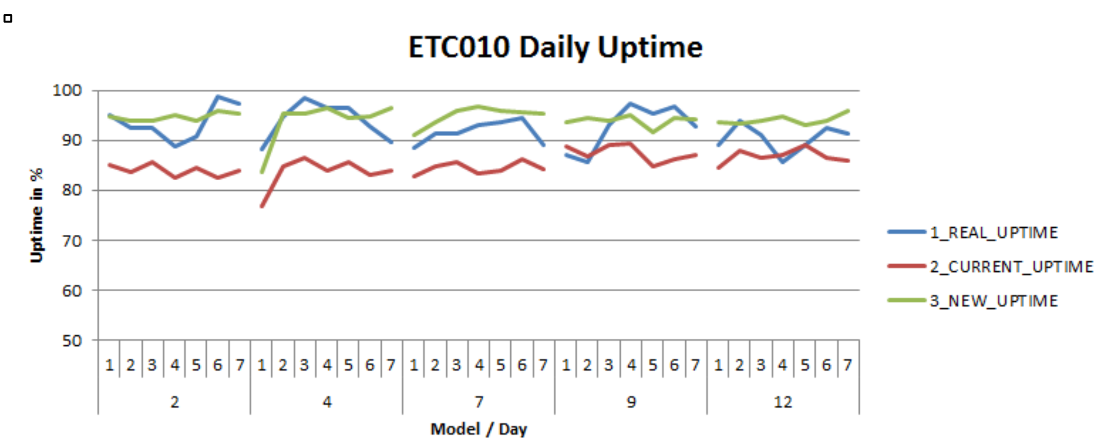


Figure 7: ETC010 daily uptime curves of real uptime (1), old approach (2), new approach (3).

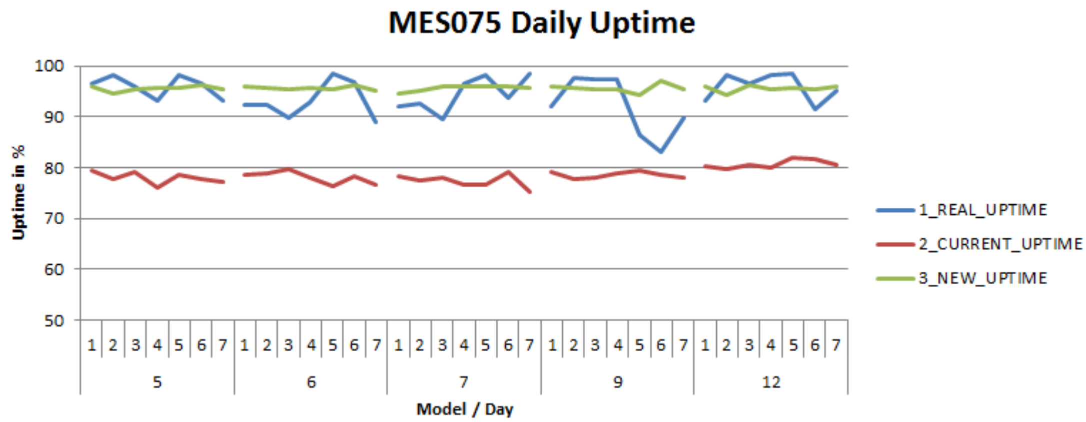


Figure 8: MES075 daily uptime curves of real uptime (1), old approach (2), new approach (3).

To evaluate the quality of the down generation algorithm, a quality assessment indicator has been defined as follows:

$$s = \frac{\sum_{e\in E}\sum_{m\in M}\sum_{d\in D}\left(U_{sim,e,m,d} - U_{real,e,m,d}\right)}{\overline{\overline{E}} * \overline{\overline{M}} * \overline{\overline{D}}}$$

For a given set of equipment groups E, models M and simulated days D the average of the daily difference between the uptime in simulation U_{sim} and the corresponding uptime in reality U_{sim} is calculated. For the three equipment groups the average uptime difference has been summarized in Table 3.

Table 3: Improvements of daily average uptime with new down time modeling approach.

Equipment Group	s (old approach)	s (new approach)
ASH001	-3.3	-0.2
ETC023	-5.8	3.0
MES075	-6.5	1.3

4 SUMMARY AND OUTLOOK

In this paper a hybrid tool down modeling approach is presented. The approach combines the methods of random and deterministic down modeling, such that forecast quality is improved by reducing the stochastic effects in short-term simulation and improving the modeling quality of the remaining stochastic elements. To achieve this, historical down events were queried, categorized and post-processed, to ensure that only relevant and typical downs are used for further analysis. Then, downs were grouped in collections and assessed to determine the best modeling approach that enhances the down modeling quality. In order to make best use of the derived down distributions, a new distribution functionality was developed in AutoSched AP. The application of this new method resulted in a better tool down time modeling as compared to the previous approach used. The down modeling method as described in this paper has been automated as part of the short-term simulation solution deployment at Infineon Dresden.

As mentioned in Section 2.3 an additional way of down modeling is triggering downs dynamically during simulation runtime. This approach is especially targeted towards counter-based PMs. These PMs are typically triggered after a defined number of wafers had been processed or a certain amount of operating hours had been accumulated on the equipment. It is possible to model these counters within the simulation, whereby the counter values are dynamically increased depending on wafer moves and processes executed. Frequent synchronization with the real equipment's counter will reduce deviation. Once a defined threshold is reached, a PM could be triggered and the counter is reset.

Unfortunately this approach also faces challenges. In a fab, maintenance work is not always started directly when the counter threshold is reached. In some cases PM are conducted earlier, for example because availability of service personnel. In other cases PM are delayed, because the equipment can still safely be operated, despite the counter value. How counter-based PM are handled varies not only from fab to fab, but also from production area to production area. Even between equipment there can be differences in handling. As these PM are typically long downs, they yield a big error if they are executed at the wrong point in time in the simulation. But when the trigger conditions are clear, the dynamic triggering of counter-based PM will further reduce stochastic effects in the short-term simulation. This is one area of work that we are currently conducting to further reduce the error in the simulation model.

REFERENCES

D-SIMLAB, 2014. "D-SIMCON – An Overview". http://www.d-simlab.com/dsimcon.htm. Accessed April. 16th, 2014

AutoSched AP, 2014. "AutoSched® AP | Applied Materials". http://www.appliedmaterials.com/global-services/automation-software/autosched-ap. Accessed May. 15th, 2014

Scholl, W., B.P. Gan, D. Noack, P. Preuss, M.L. Peh, P. Lendermann, and O. Rose. 2010. "Towards Realization of a High-Fidelity Simulation Model for Short-Term Horizon Forecasting in Wafer Fabrication Facilities." In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 2563-2574. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Law, A. M.. 2007. *Simulation Modeling and Analysis.* 4. ed. Boston: McGraw-Hill, Int.

Achermann, D. 2008. *Modelling, Simulation and Optimization of Maintenance Strategies under Consideration of Logistic Processes.* Swiss Federal Institute of Technology Zurich. Diss. ETH No. 18152.

AUTHOR BIOGRAPHIES

PATRICK PREUSS is a Project Manager and the Deputy Manager Germany Operations of D-SIMLAB Technologies (Germany). He has been working in the development of simulation-based applications for Airbus, German Aerospace Centre and Infineon with focus on data analysis and heuristic optimization methods since 2005. Patrick holds a M.S. degree in computer science from Dresden University of Technology. His email address is patrick@d-simlab.com.

ANDRÉ NAUMANN is a Software Engineer at D-SIMLAB Technologies (Germany). He studied computer science with focus on simulation-based optimization at Dresden University of Technology and graduated with a diploma degree. Since 2012 André is working D-SIMLAB Technologies (Germany) in deployment projects for customers in semiconductor industry. His email address is andre@d-simlab.com.

WOLFGANG SCHOLL works as a Senior Staff Expert for modeling and simulation for Infineon Technologies in Dresden (Germany). He has studied physics at the Technical University of Chemnitz (Germany) and graduated in solid-state physics in 1984. From 1984 to 1995 he worked as a process engineer for ZMD in Dresden. In 1996 he joined Infineon Technologies (former SIMEC) and worked on the field of capacity planning. Since 2003 he is responsible for fab simulation. His email address is wolfgang.scholl@infineon.com.

BOON PING GAN is the CTO of D-SIMLAB Technologies (Singapore). He has been involved in simulation technology application and development since 1995, with primary focus on developing parallel and distributed simulation technology for complex systems such as semiconductor manufacturing and aviation spare inventory management. He was also responsible for several operations improvement projects with wafer fabrication clients which concluded with multi-million dollar savings. He is now responsible for managing the technology and product development at D-SIMLAB as well as execution of projects in the semiconductor and engineering domains. He holds a Master of Applied Science degree, specializing in Computer Engineering. His email address is boonping@d-simlab.com.

PETER LENDERMANN is the Co-Founder and CEO of D-SIMLAB Technologies, a Singapore-based company providing simulation-based decision support solutions and services to Aerospace, Semiconductor Manufacturing and other asset-intensive industries. Prior to this he worked at the Singapore Institute of Manufacturing Technology where he led the simulation-related research activities until spinning them off into D-SIMLAB Technologies. He has been engaged in the simulation community since the early 1990's. Peter holds a PhD in Applied High-Energy Physics from Humboldt University in Berlin (Germany) and an MBA in International Economics and Management from SDA Bocconi in Milan (Italy). His email address is peter@d-simlab.com.