Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

HARMONIZING OPERATIONS MANAGEMENT OF KEY STAKEHOLDERS IN WAFER FAB USING DISCRETE EVENT SIMULATION

Georg Seidel

Infineon Technologies Austria AG
Siemensstraße 2
Villach, 9500, AUSTRIA

Ching Foong Lee
Ai Mei Kam

Infineon Technologies (Kulim) Sdn Bhd
Jalan Hi-Tech 7, Industrial Zone Phase II,
Hi-Tech Park,
Kulim, Kedah 09000, MALAYSIA

Boon Ping Gan
Chew Wye Chan

D-SIMLAB Technologies Pte Ltd
8 Jurong Town Hall Road #23-05
JTC Summit
Singapore 609434, SINGAPORE

Andre Naumann
Patrick Preuss

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

ABSTRACT

Operations meeting in a wafer fab involves daily alignment of action items among key stakeholders: operations, maintenance, engineering and planning department. They have conflicting job functions. The maintenance department is required to conduct preventive maintenance (PM) to improve tools' reliability The engineering department is required to qualify new products and processes. Both cases interfere with the flow of production lots. The planning department must ensure production ramp up. This can have a short term impact on overall fab delivery and capacity. The primary challenge is to reach aligned decisions e.g. the best timing for PM or optimizing dispatch prioritization of production and development lots to ensure on-time delivery while maximizing capacity and tool utilization. In this paper we discuss the associated modelling issues of a 7-day simulation-based forecast, providing forecast of incoming WIP, moves and utilization at work center level. The simulation forecast consistently achieved an accuracy above 90%.

1 INTRODUCTION

Managing daily operations in a front-end wafer fab is a complex task. Different departments are involved and they have often different targets and key figures. For example line control is, among others, typically responsible for lot cycle time and on time delivery. . Process engineers are usually more concerned about process stability and yield. Maintenance engineers are in charge of tool stability and availability. All these departments have to deal with conflicting targets. Line control should try to maximize loading, throughput and tool utilization and at the same time minimize lot cycle time and provide delivery date reliability. Maintenance engineers will try to maximize tool availability and minimize costs (e.g. personal and spare parts).

These conflicting targets lead automatically to difficult decision making. Alignment meetings are in place to facilitate the decision process at Infineon's production site in Kulim (Malaysia). However a lot of decisions are difficult to make due to the missing knowledge of how the decisions will impact key

performance figures. One specific example for such decision making is the planning of PM (preventive maintenance).

Maintenance engineers have to take care of PM deadlines, appropriate staffing to conduct PM and the availability of spare parts. Line control is interested in minimal disturbances of line performance caused by any PM. Until recently there has been no well-defined business process available to decide when to conduct the PM.

Another specific example for a conflicting target is tool reservation for engineering. Engineering department requires equipment time for testing and developing new processes. Line control, as mentioned earlier, does not appreciate any significant disturbances of line performance due to e.g. missing tool availability. Fab simulation can facilitate the decision making in both cases by providing a reliable WIP flow forecast. Similar works were done to forecast lot arrival for defect density area for the planning of resource requirement and adjusting the sampling rate (Scholl et. al. 2011).

2 CRITICIAL MODELLING CONSIDERATIONS FOR REALIZATION OF SHORT TERM SIMULATION

In order to provide a forecast to facilitate the decision making process among departments, a simulation model is built on *D-SIMCON Forecaster* (D-SIMLAB Technologies 2017) to provide 7-day forecast for all work center. The simulation runs daily automatically at night. Report charts are generated after the run and are available the next morning. Key figures provided within reports are daily incoming WIP, moves, average WIP level and utilization for each work center. In this section we discuss the key modelling elements that ensure an aggregated forecast accuracy of 90% or higher. These key modelling elements are: process flows, dedication, wafer start, equipment model, dispatch rules and equipment uptime. Similar works were done by S. Bagchi and team to model a 300mm wafer fab (Bagchi et. al. 2008) where the model generation was fully automated.

2.1 Process Flows

A process flow is a sequence of processing steps that a lot has to undergo to finish overall processing. The important modelling elements that have been considered for each manufacturing step are: recipe, sampling, split-merge and queue time constraint. A specific recipe determines the needed process for a certain process step. Each equipment is capable to run a set of recipes. The recipe-equipment combination determines the process time needed for a process step and thus models the equipment throughput, lot cycle time and equipment capacity constraint.

Sampling is used to model measurement and inspection steps with the sampling rate derived from three months of historical data. It is essential to model these steps to ensure that lot cycle times are accurately forecasted.

The split-merge function at a process step is used to model pilot runs for certain lithography and chemical-mechanical-polishing processes. Instead of modelling all details to decide when to make a pilot run, average split rates and split sizes for specific process steps have been calculated from historical data. This simplification reduces modelling complexity without big impact on forecast accuracy.

Typically some segments of a process flow require that a lot completes all process steps in the segment within a certain time limit. These steps are called queue time constraint steps. Production lots are held back at the beginning of the queue time constraint steps until it is certain that the lots can flow through the queue time steps within the time limit. It is thus essential to model the queue time constraint steps with its associated control mechanism (how to hold lots back) to ensure that the WIP forecast at these steps is accurate.

2.2 Dedication

Each equipment is dedicated to run a set of recipes, hence, a set of steps in the process flow. This imposes capacity limits for each recipe. Additionally, short term blocking is possible for specific products or routes on certain equipment due to temporary process constraints. The simulation model is initialized with existing short term blockings at the moment of model generation and it is assumed that the blocking stays valid until end of simulation time.

2.3 Wafer Start

The process of releasing lots of different products to the fab is called "Wafer Start". The planning process behind "Wafer Start" is complicated because capacity constraints and customer deadlines have to be considered. In order to provide an accurate forecast, the simulation model was fed with the actual wafer start plan. However, there is a practical limitation on the availability of the wafer start plan. Fabs typically do not plan in detail more than one week in advance. As such, the wafer start plan is replicated for the days without any detailed plan. This is a reasonable assumption because product mix does not change significantly from one week to another. At times production department starts lots earlier or later than planned. In these cases early started lots are removed from wafer start plan while delayed lots are released as planned.

2.4 Equipment Model

Equipment in the fab can generally be categorized into five different tool types: (i) Diffusion, (ii) Thin-Film, (iii) Etch (CVD/PVD/SPU), (iv) Lithography, and (v) Metrology. These equipment types differ in the way they process wafers. They can process one production lot as a whole or one/multiple wafer(s) concurrently or a collection of lots as a batch. How process time and throughput are simulated depends on the tool type.. Equipment are first categorized based on this criteria.

In modern wafer fabs, there are many cluster tools with multiple chambers that can process wafers from the same lot or from different lots concurrently. Cluster tools can be configured to run in either serial or parallel mode. What is the right abstraction level to model cluster tool behavior? One option is using a detailed model where each wafer move going through process chambers is modelled. Another option is the application of abstracted models where no chambers are modelled at all. Every approach has its shortcomings. Detailed modelling leads to high execution time. Omitting chamber modelling means that any chamber unavailability that changes equipment throughput will not be considered. Our solution is to model the cluster tools using a chamber matrix which defines the chamber set, required to run a particular recipe. Moreover we use a corresponding throughput reduction factor if chambers are down.

2.5 Dispatch Rule

Dispatch rules determine in which sequence lots are processed by each equipment. Not modelling dispatch rules in simulation could result in a different simulation WIP profile compared to the production line. Therefore we model all global dispatch rules. Some examples are: lot priority, lot due date, queue time constraint, setup avoidance, etc. These dispatch rules have an impact on more than 85% of the WIP. Some are very complex and have only minor impact wherefore they have been omitted to reduce model complexity.

2.6 Equipment Uptime

To model equipment uptime we use two different down events: scheduled and unscheduled downs. In our simulation model scheduled down times are modelled as deterministic events. Planned start date and estimated duration for the scheduled down must be available. Unscheduled downs, however, are modelled as random events using a statistical distribution that is derived from three months historical data.

2.7 Model Initialization: Current WIP and Equipment Status

The simulation model is initialized with the current state of the fab. It includes the current positions of all lots, their associated statuses (processing, queueing, on-hold, split or in rework) and the current equipment statuses (such as processing, stand-by, scheduled down, unscheduled down, engineering or non-scheduled). For equipment that is down at the beginning of simulation, the down duration is estimated from historical data.

3 MODEL VALIDATION APPROACH

A simulation model is built based on the key modelling elements discussed in Section 2. Model validation is the most important aspect of any simulation project. Model validation involves comparison of key figures generated by simulation with key figures observed in reality. We need to ensure that equipment capacity, work center capacity and lot cycle time behavior are a close representation of the reality. Prerequisite for validation is to define a method of how to measure forecast quality. It is challenging to determine if the work center or the model itself is valid without such a measurement method. Furthermore, a permanent monitoring of forecast quality even after validation and deployment is required. Only then deviations between real fab and model behavior can be detected, enabling the engineers to conduct counter actions and making sure that the forecast quality is consistently high. This ensures that any decisions derived from the forecast are effective and relevant.

The forecast quality is measured by calculating deviations between simulation and reality. It is defined as gap of a KPI (key performance indicator, e.g. incoming WIP, moves, WIP, etc.) at a specific entity (equipment, work center, etc.) within a defined time frame (day, week, month, etc.) of the simulation period. This gap can range from -1 to 1. A gap value of 0 indicates an exact match of simulation and reality. Describing the forecast quality as a gap allows us to analyze the direction of deviation, whether the KPI simulation value exceeds the real value or vice versa (Figure 1).

$$Gap_{k,m,e,t} = \begin{cases} \dfrac{sim_{k,m,e,t} - real_{k,m,e,t}}{\max(sim_{k,m,e,t}, real_{k,m,e,t})}, & \max(sim_{k,m,e,t}, real_{k,m,e,t}) \neq 0 \\ 0, & \max(sim_{k,m,e,t}, real_{k,m,e,t}) = 0 \end{cases}$$

Figure 1: Gap of KPI k at entity e within time period t in model m.

$$Gap_{k,m} = \frac{\sum_e W_e * \dfrac{\sum_t W_t * Gap_{k,m,e,t}}{\sum_t W_t}}{\sum_e W_e}$$

Figure 2: Weighted gap of KPI k in model m.

The gap values are aggregated to measure overall forecast quality as shown in Figure 2. Examples for gap values are the average move gap of a work center throughout the entire simulation period or alternatively the WIP gap of the entire model within a particular day during the simulation period. In addition, these averages can be weighted to provide flexibility of emphasizing certain entities such as mission critical equipment or certain time periods.

Target values were set for model validation after we defined our measurement method. Due to re-entrant WIP flow, a typical wafer fab characteristic, this validation process was challenging. Discrepancies in one work center might propagate to the following downstream groups. This behavior might mask existing modelling issues or create issues where there are none (Figure 3). Analyzing the

upstream groups to find the root cause of a problem might end-up at the starting point due to the circular nature of the wafer fab WIP flow. To solve this problem we developed the Model Validation Engine (MVE). Two phases are considered: the real workload phase and the integrated phase.

In the first phase we validated individual work center behavior for KPIs such as moves, WIP and tool utilization. Instead of free WIP flow (lots moving from work center to work center following the process flow), each work center was fed with an actual incoming workload gathered from real trace data. All lots are discarded from the model once they are processed at the work center. Thus, any downstream group will not be fed with the workload from its predecessor. Only the actual incoming workload will be considered in this phase. Any KPI discrepancy between simulation and reality for a specific work center can only be due to modelling issues at this work center (Figure 4). Work centers can be analyzed and corrected without interfering with each other using this approach. Another advantage is that multiple work centers can be validated at the same time by different engineers. This can speed up model validation dramatically.

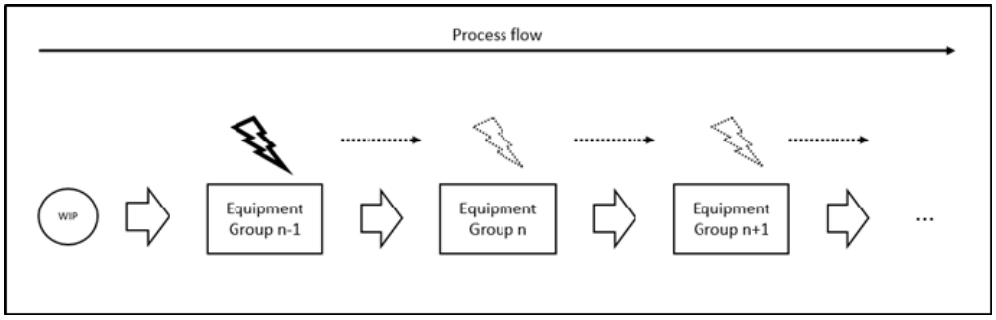

Figure 3: Conventional lot flow with error propagation.

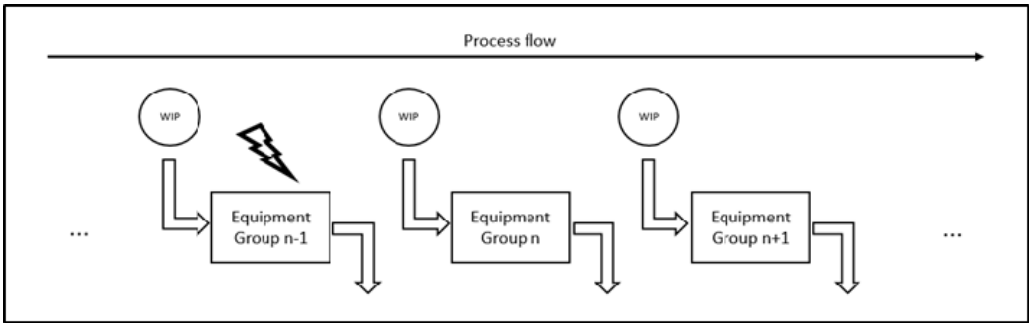

Figure 4: MVE lot flow with isolated work centers.

After each individual work center behavior is optimized and forecast quality meets its target, we start the second phase of validation. In this phase the work center models will be merged and the WIP flows through the model following the process flows. It is necessary to validate the modelling elements associated with the interactions of work centers, such as e.g. time coupling rules. We focused on validating the most important key figures like incoming WIP, moves and average WIP. Forecast qualities of each work center and the model in general were compared with their targets. When targets were met, it was concluded that the model was validated and could be used for decision making. Using this approach we managed to validate the model and achieved a forecast accuracy of above 90% within three months.

4 USE CASES

Figure 5 illustrates the usage of simulation forecast for alignment of operational decisions among maintenance, production, planning and engineering departments. Two examples for this decisions are: (i) selection of the best day to conduct preventive maintenance (PM) and to facilitate maintenance resource planning, (ii) selection of the best day to conduct engineering activities. In both cases, the objective is to minimize line interference. To conduct PM and engineering related activities, the departments should therefore select days with low WIP and low production utilization.

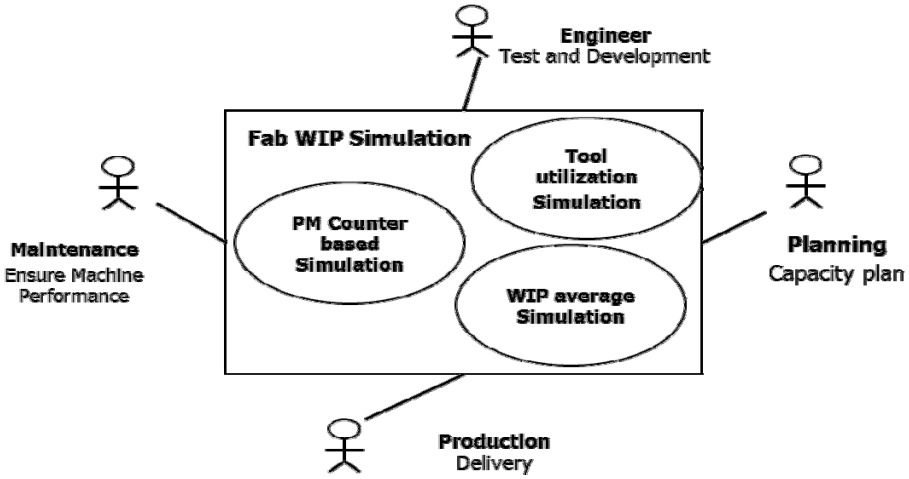

Figure 5: Diagram shows key stakeholder for alignment processes.

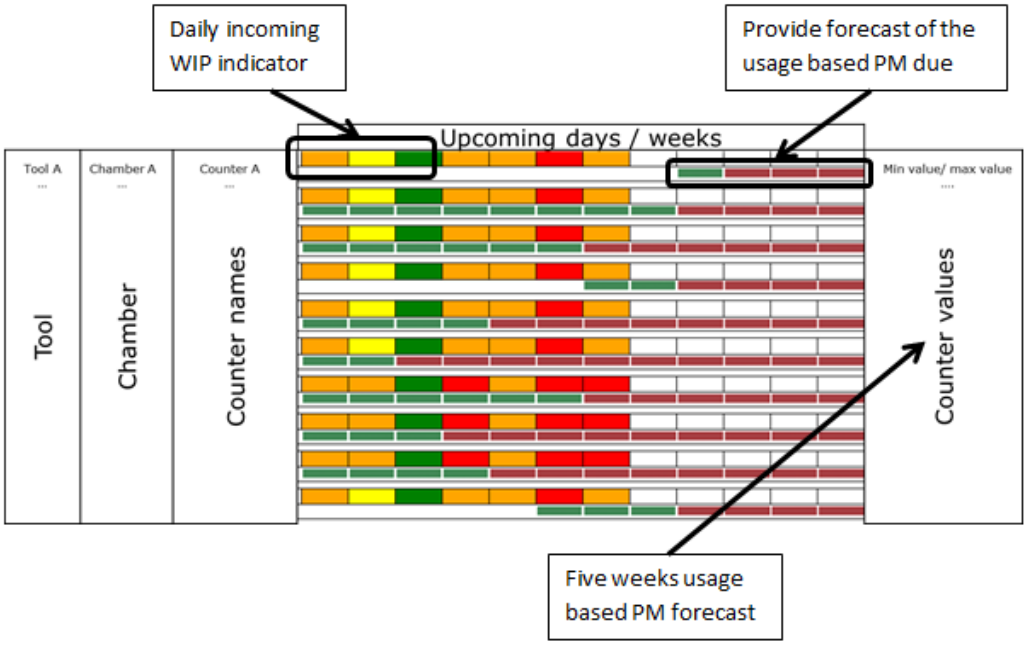

Figure 6: Preventive Maintenance report.

Forecasts are presented as web-based reports (generated daily) as illustrated in Figure 6. These reports are then used for decision making in alignment meetings. Each row of rectangles in Figure 6 represents

the expected incoming WIP level for an equipment. The color of the rectangle indicates the volume of incoming WIP in reference to threshold values, which were derived from capacity limits of the equipment. Green color represents low, yellow represents medium and red represents high incoming WIP. Days with green rectangles are best suited for PM activities.

Some equipment in the fab requires PM to be conducted based on its past usage. Simulation is used to provide a forecast of the PM due date. This is represented by thin rectangles below the bigger rectangles. A green color of the thin rectangle indicates that a minimum threshold value will be reached. Red color indicates that the equipment must be shut down for PM because the upper threshold value will be reached. Maintenance manager and line control experts are using this report to align on PM time. To facilitate longer time horizon planning the simulation forecast also provides a five-weeks look ahead as illustrated at the right side of the chart by interpolating the average moves made in the seven-days forecast.

Utilization reports (see Figure 7) are used to determine the best day to conduct engineering related activities. Each stack bar represents the daily forecast of tool utilization (PR) in green, standby (SB) in yellow, scheduled down (SD) in purple, and unscheduled down (UD) in red. Obviously, days with lower expected utilization (such as the last three days of the forecast) are more suitable for conducting engineering activities as compared to the days with higher expected utilization (such as the first 4 days). To ensure user confidence we show the utilization comparison between simulation and reality for the past three days (three left most set of stack bars).

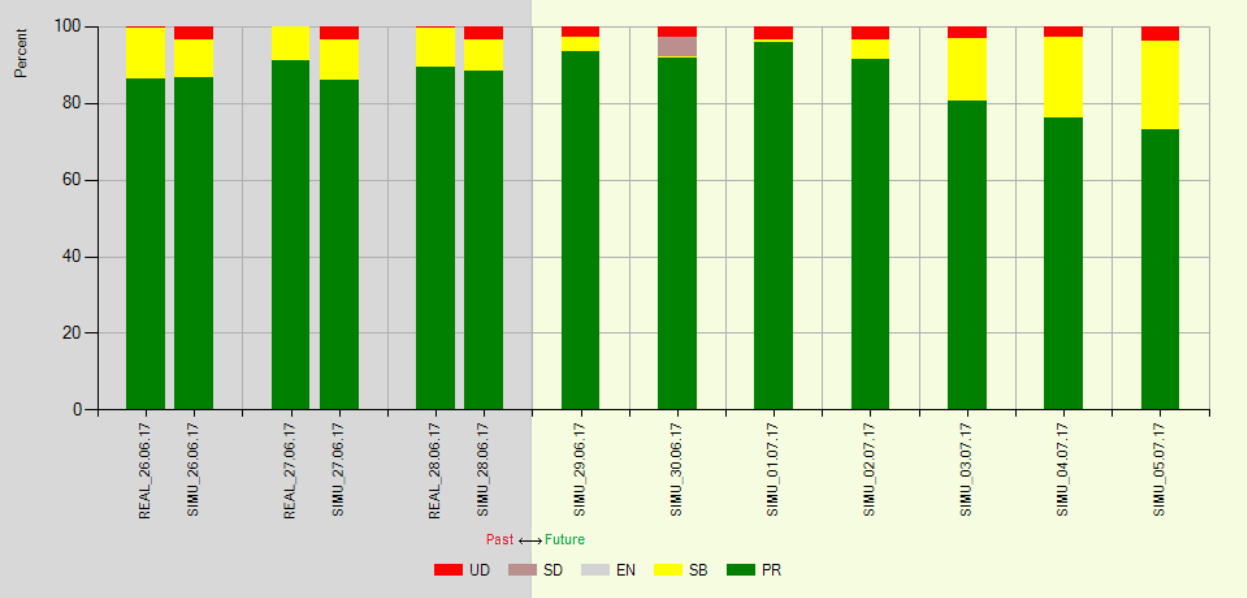

Figure 7: Work center utilization chart.

5 CONCLUSIONS

Short term simulation has helped operations to facilitate decision making in a more formal way. PM planning and engineering activities planning are the first use cases in place. Further use cases such as workload forecast for operator resource planning, hot lot journey through the fab to ensure hot lots complete their expected steps in time and mask layer per week to ensure target is met are in preparation. Additional to that, discussions are ongoing to extend the simulation period for even more use cases. Examples are dispatch rules evaluation, wafer start (changes in product mix) impact on line performance and early warning system for WIP increase within next three months.

REFERENCES

Scholl, W., D. Noack, O. Rose, B.P. Gan, P. Lendermann, P. Preuss, and F.S. Pappert. 2011 "Implementation of a Simulation-based Short-term Lot Arrival Forecast in a Mature 200mm Semiconductor Fab". In *Proceedings of the 2011 Winter Simulation Conference*, edit by S. Jan, R. Creasey, J. Himmelspach, K.P. White, and M.C. Fu, 1932-1943. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Bagchi, S., C.H. Chen, S.T. Shikalgar, and M. Toner. 2008. "A Full-factory Simulator as a Daily Decision-Support Tool for 300mm Wafer Fabrication Productivity". In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. Mason, R. Hill, L. Mönch. 2021-2029. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

D-SIMLAB Technologies. 2017. Forecaster and Scenario Manager. http://www.d-simlab.com/category/d-simcon/products-d-simcon/forecaster-and-scenario-manager.

AUTHOR BIOGRAPHIES

GEORG SEIDEL is Senior Staff Engineer of Infineon Technologies Austria AG (Villach, Austria). He has been involved in simulation, WIP flow management and Industrial Engineering topics since 2000. He was responsible for WIP flow management, especially for Lot dispatching at Infineon's site in Kulim (Malaysia) from 2012 until 2015. He is now responsible to rollout Fab Simulation in Kulim and Villach. He holds a Master degree of Technical Mathematics. His email address is georg.seidel@infineon.com.

CHING FOONG LEE is Senior Specialist Engineer of Infineon Technologies (Kulim) Sdn. Bhd. She has been involved in Semiconductor System Development and Datamining since 2004. She joined Infineon Technologies Kulim in 2010 driving various projects in Production System Setup, Reporting and System Improvement under Factory Integration department. Currently she is responsible in Kulim for Simulation and WIP flow management topics under Operation Research and Engineering department. She holds Master of Business Administrative(MBA) and Bachelor Degree of Information Technology, majoring in Software Engineering. Her email address is chingfoong.lee@infineon.com.

AI MEI KAM is the Director of Infineon Technologies (Kulim) Sdn. Bhd. She has been in the semiconductor industry since 2000, with a wide range of trained knowledge in multiple areas of wafer fabrication manufacturing including operations, procurement, planning and controlling. She has been leading the site Project Office to drive strategic program to support the company objectives and targets. She has proven ability to drive organization in achieving yearly performance goals. She is now also responsible for site productivity management and heading the Operation Research Engineering team. She holds a Bachelor of Engineering degree and her e-mail address is AiMei.Kam@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 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.

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.

CHEW WYE CHAN is a Software Engineer of D-SIMLAB Technologies (Singapore). He has been working in the development of simulation-based applications for logistics and semiconductor industry with focus on simulation-based optimization since 2006. Chew Wye holds a Master of Computing degree from National University of Singapore. His email address is chew.wye@d-simlab.com.

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.