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

AN EASY APPROACH TO EXTENDING A SHORT TERM SIMULATION MODEL FOR LONG TERM FORECAST IN SEMICONDUCTOR INDUSTRY

Marcin Mosinski
Tobias Weissgaerber

Robert Bosch GmbH
Postfach 13 42
Reutlingen, Baden-Württemberg 72703,
GERMANY

Soo Leen Low
Boon Ping Gan

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

Patrick Preuss

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

ABSTRACT

The operational decision making in the BOSCH's 200mm wafer fabrication facility has been guided by short term simulation forecasts. The forecasts provides the capability of identifying daily bottlenecks, forecasting daily fab outs, optimizing the preventive maintenance plans and personal resource planning. Now there is a pressing need to extend the forecast time horizon to several months for making decisions such as analyzing different ramp up scenarios, evaluating the impact of dispatch rules, identifying bottlenecks for capital investment, etc. As the short term model has achieved forecast accuracy of above 90%, it is used as the basis to generate the long term model. In this paper, we discuss the key issues associated with this model generation process. These issues are: process flows compression, flexible equipment dedications, model warm-up, wafer start generation, and future fab capacity changes. Our approach enables us to use the same model generation framework for both models.

1 INTRODUCTION

Discrete event simulation has advanced the production control process in the wafer fabrication plants, primarily used as a short and long term decision support tool. In this paper, a short term decision support refers to a time horizon of less than a month while a long term decision support refers to a time horizon of at least 12 months. Some examples of the decision support include forecasting incoming WIP at equipment group level to ensure better preventive maintenance (PM) planning and sampling rate adjustments (Scholl et. al. 2010, Scholl et. al. 2012), lot movement forecast to ensure that equipment is easily qualified to run associated processes when the lot arrives (Scholl et. al. 2016), and study the effect of hot lots on factory key performance indictors (Narahari and Khan 1997).

In the past years Bosch Reutlingen has embarked on the journey of introducing a 35-day WIP forecast, through short term simulation using *D-SIMCON Forecaster* (D-SIMLAB 2017). This is used to address various operational challenges in daily operation meetings, such as: (i) an early warning system to identify bottlenecks for the next days to secure operator resource planning, optimizing the PM plans by avoiding days with high incoming WIP, (ii) fab outs forecast to provide visibility into tardiness of customer critical lots so that corrective action can be taken if lots are going to miss delivery commitment

978-1-5386-3428-8/17/$31.00 ©2017 IEEE

dates, (iii) fab outs forecast for detailed planning at the wafer test area, (iv) target setting for all work areas, and (v) a better understanding of line dynamics. The simulation-based forecast has helped to improve the discussion among line control, maintenance, and planning department as decisions are now based upon numbers derived from a widely accepted methodology rather than simply the experience of individuals.

Due to the success in the application of short term simulation, the confidence in the use of simulation has increased. Thus, Bosch has decided to extend the time horizon of the simulation run to address questions such as the impact of different product mixes to line performance and associated capital investment, the impact of enhanced dispatch rules to fab productivity and utilization, the impact of weekly wafer start plan to line balance, at times evaluate varying recovery strategies from a production line incident, and use as a bridge between production, line control, finance, and management. In this paper, we refer to the extended time horizon simulation as long term simulation. The future vision of this solution is to reach a state of holistic value stream management from wafer to chip manufacturing, more precisely connecting the wafer fabrication, MEMS, and assembly & test models for a cross site forecasting. Similar works have been completed to model cross fab capacity sharing in Gan et. al. (2004). As the short term simulation solution was successfully rolled out, the most logical step to implement long term simulation is reusing the short term simulation framework.

This paper begins with the foundation of discussion by describing the current short term simulation framework in Section 2. This is followed by a discussion on issues associated with the extension of the time horizon of short term simulation model in Section 3. In Section 4 an experimental study on the impact of simplifying process flows to fab KPIs is described as well as a case study to illustrate the use of long term simulation model for dispatch rules enhancement. This paper is concluded in Section 5 with a focus on the future work around the application of simulation.

2 SHORT TERM SIMULATION FRAMEWORK

The short term simulation forecast is run daily and provides a forecast for the next 35 days. The forecast consists of daily key performance indicator (KPI) numbers for equipment groups, products, and the fab, as summarized in Table 1. Due to the high frequency run of the forecast, the simulation model has been generated automatically by consolidating data from live production databases (see Figure 1). This ensures that any simulation forecast is not outdated by the time the model is created. In order to achieve high forecast accuracy (and thus acceptance of the solution), the forecast has to be based upon the current line situation, upcoming wafer start and PM plans and a complete set of process flows (instead of representative ones). A snapshot of the line situation is taken to jump start the simulation model. It includes the current WIP (all lots in the production line including their status), the current equipment state (productive, scheduled/unscheduled down, standby, or non-schedule) and the temporary dedication blocking at all equipment. The validity of the model is continuously monitored by comparing the KPIs observed in reality versus the forecasted KPIs. Any sudden drop in forecast quality triggers the simulation specialist to investigate if the deteriorating quality is due to short term random events or data modelling issues. This process ensures that any decision derived from the forecast is always valid. Table 2 summarizes the functionalities of key components of the *D-SIMCON Forecaster*.

Table 1: Forecast key performance indictor.

Element	Key Performance Indictor (KPI)
Equipment Group	Incoming WIP, moves (number of wafers completed), average WIP, average queue time and utilization.
Product	Number of wafers started, number of wafers completed, and average cycle time.
Fab	Average WIP, total moves (number of wafers moving from one step to another), average cycle time.

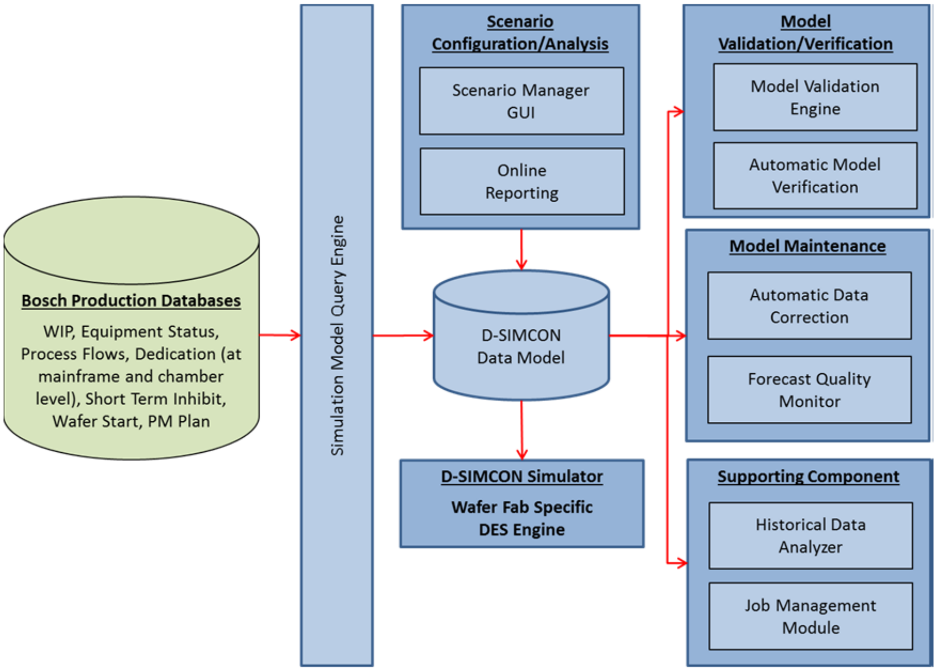

Bosch Production Databases
WIP, Equipment Status, Process Flows, Dedication (at mainframe and chamber level), Short Term Inhibit, Wafer Start, PM Plan
Simulation Model Query Engine
Scenario Configuration/Analysis
Scenario Manager GUI
Online Reporting
D-SIMCON Data Model
D-SIMCON Simulator
Wafer Fab Specific DES Engine
Model Validation/Verification
Model Validation Engine
Automatic Model Verification
Model Maintenance
Automatic Data Correction
Forecast Quality Monitor
Supporting Component
Historical Data Analyzer
Job Management Module

Figure 1: *D-SIMCON Forecaster* architecture.

Table 2: *D-SIMCON Forecaster* component functionalities.

D-SIMCON Components	**Functionalities**
Automatic Model Verification	Comparing the key modelling attributes between simulation and reality as well as highlight gaps.
Model Validation Engine	Feeding simulation model with historical lot trace information to validate the behaviour of each equipment group; with additional functionalities of replacing data for some modelling elements (such as wafer start, PM, UD) with real data to validate the model by “reducing” random/inaccurate elements
Automatic Data Correction	Enable user to configure to allow the system to automatically replace data with big gaps, and the frequency of conducting such a check
Forecast Quality Monitor	Assists in the model validation/verification exercise to measure the forecast gap between simulation and reality
Historical Data Analyser	Derive sampling rate, rework rate, split-merge rate/size, uptime distribution of equipment group/equipment
Job Management Module	To manage replication runs, and consolidate statistics
Online Reporting	Web based reporting for scenario analysis (require customization)
Scenario Manager GUI	Web based GUI to view and edit scenarios, with extended functionalities of defining actions to add/remove tools, alter process efficiency, and wafer starts (requires customization)

3 ISSUES ASSOCIATED WITH EXTENDING SHORT TERM SIMULATION FOR LONG TERM SIMULATION

In this section a few key modelling elements that require changes or modifications to extend short term simulation into long term simulation studies are discussed. The objective is to attain the extension in an easy manner without any user involvement or interaction.

3.1 Process Flows

With very high level detail of the short term model, all process flows are considered in the model. In the long term model, the planning granularity is usually at the product group level. Partial reason is that it is usually challenging to obtain the volume of each of the products for a longer time period. In addition changes in technology may introduce a new product to the fab, where a complete process flow is still not yet available at the time of study. Therefore it is more practical to select representative process flows for each planned product group instead of considering complete sets of process flows.

There are a few criteria to be considered when selecting representative process flows. Firstly, the process flows must have sufficient historical data to derive random event inputs such as sampling rate, rework rate, and hold rate/duration. Secondly, process flows within a product group need to be categorized based on similarity in terms of number of stages, recipes and layers. Each product group is then represented by one process flow from each category. The purpose is to ensure that the selected representative process flows are good candidates to portray the fab performance trend in the long term model. Another simplification for the process flow is achieved by dropping steps that are not essential to be modelled. Some of the data collection steps were dropped from the simulation model as they do not have any capacity impact but only cycle time delay. One example of such a delay step is the lot hold step where there is no resources required to process the lot but a placeholder to collect data for lot hold. These steps are represented as delay steps with unlimited capacity, with the delay time derived from historical data analysis.

3.2 Wafer Starts Plan

The short term model is fed with a daily lot-by-lot wafer start plan. This model is not feasible when it comes to the long term simulation as future wafer starts are typically planned at the weekly or monthly level. Therefore the generation of a wafer start plan from the provided weekly or monthly product group volume is required. There are two possibilities to achieve this: (i) define a wafer starts rate for each of the products, or (ii) generate the wafer start plan based on a scheduling algorithm. Wafer starts rate is the more straight forward way to be used in the long term model. This approach implies that wafers are released in a constant and systematic way. This is not likely to happen in production reality.

As a result we have decided on the second approach. The algorithm considers the total volume to be started per week or month, and forms batches based on the first six furnace steps. It then assigns a start date and time for each batch. Daily volume is kept at a constant to ensure that there is no over starting of wafers which would result in WIP waves. One key consideration is that it might not be feasible to form batches for some low volume products. For these products, the algorithm will fit the volume randomly to days within a week. Each started lot is assigned a due date based on the start date plus the target cycle time. The due date is essential as dispatch rules use this value to drive the lot moves through the fab. The algorithm also allows the user to provide lot priority as typically a fab is run with some percentage of high priority lots.

3.3 Model Warm-up

In the short term model, the simulation is initialized with the snapshot WIP and equipment state. As long term simulation might involve simulation study that is in the future, the WIP mix profile might not be the same as of today due to rapid changes in product technology. Therefor driving a model with today’s WIP might not be a good representation. To provide flexibility in long term simulation three different options are provided as the starting point of the simulation run: (i) the use of current WIP if the forecast starts from today, (ii) the use of current WIP and continues into the future if the time period is in the future but the WIP mix has not changed, or (iii) the generation of an artificial WIP based on the wafer start profile provided if the WIP mix has changed. The artificial WIP algorithm provides multiple options by considering the flow factor at each process steps.

3.4 Equipment Dedication

Equipment dedication is specified at the mainframe and chamber level. Short term temporary blocking of lots belonging to a specific product or specific step is considered in the short term simulations due to its capacity impact or limitation, which must be captured to ensure good forecast accuracy. In addition, alternative dedication is driven based on an alternative rate that is derived from historical lot trace analysis. However, full flexibility on equipment dedication is required in the long term model as the study is to assume the fab is running under an ideal situation. The alternative dedication cannot be modeled as an alternative rate because future product mixes might require a different capacity profiles. All alternative dedications are considered as equal in the case of the long term model.

3.5 Tool Availability

The short term model mainly uses a deterministic modelling approach and stochastic behavior is always kept at a minimum (Preuss et. al. 2014). The PM plan where the information is typically already available at the time of simulation is fed as an input to the short term simulation model. The only randomness that is introduced is the unscheduled down events, which is modelled as random events based on statistical distribution. For the long term model, PM data is not available. As such, PMs have to be modelled as random events too. For the purpose of our model, the availability is modelled with a single uptime distribution for each equipment. The uptime distribution is defined with a mean-time-to-failure and a mean-time-to-repair parameters. The distribution used is derived from historical data analysis. As opposed to short term simulations, the time horizon used to derive the statistical distribution is one year instead of three months. The reason being is that the short term uptime effect of a tool is not supposed to be considered in the long term model, but it is important for the short term model.

3.6 Future Tool Requirement

One last important consideration in a long term model is the future introduction of new equipment into the simulation model due to capacity expansion. This is typically not required in a short term simulation model as all equipment that is qualified for the production should already exist in the production database. To handle this level of complexity we designed a Graphical User Interface that allows the user to add a new equipment by selecting an existing equipment as a blueprint for its behavior. The convenience for users is that they do not need to be concerned with the complexity of adding equipment involving assigning the equipment to an equipment group, assigning recipe dedication, defining uptime behavior, etc.

4 RESULTS AND ANALYSIS

4.1 Model Validation

To ensure the usability of the long term model we conducted a model validation exercise using a 5-week time period. We fed the long term model with a wafer start plan that was a replica of the first week wafer start plan, and compared the equipment groups incoming WIP levels and moves between simulation and reality. Figure 2 and Figure 3 below show the validation results. The forecast quality percentage is calculated using the following formula:

$$KPI\ Forecast\ Quality\ (\%) = 1 - \frac{|Real\ KPI - Simulation\ KPI|}{MAX(Real\ KPI, Simulation\ KPI)},$$

where the KPI is incoming WIP and Moves for the comparison in Figure 2 and Figure 3 respectively. It can be seen that there was a close match (above 90% forecast quality for first 3 weeks of simulation) between simulation and reality for the 7 example equipment group KPIs. These equipment groups are chosen as they are the most critical bottleneck tools in the fab. The gap increased as the time horizon extended beyond the fourth week. This growing gap is primarily explained by WIP built-up at some critical bottlenecks as the wafer start at the start of the simulation week was higher than expected. The fab capacity is only able to handle a momentarily increase in wafer starts and must be compensated with a lower starts in the subsequent weeks. In addition, the forecast quality is also expected to drop as we forecast further into the future due to the cumulative effect of the random events.

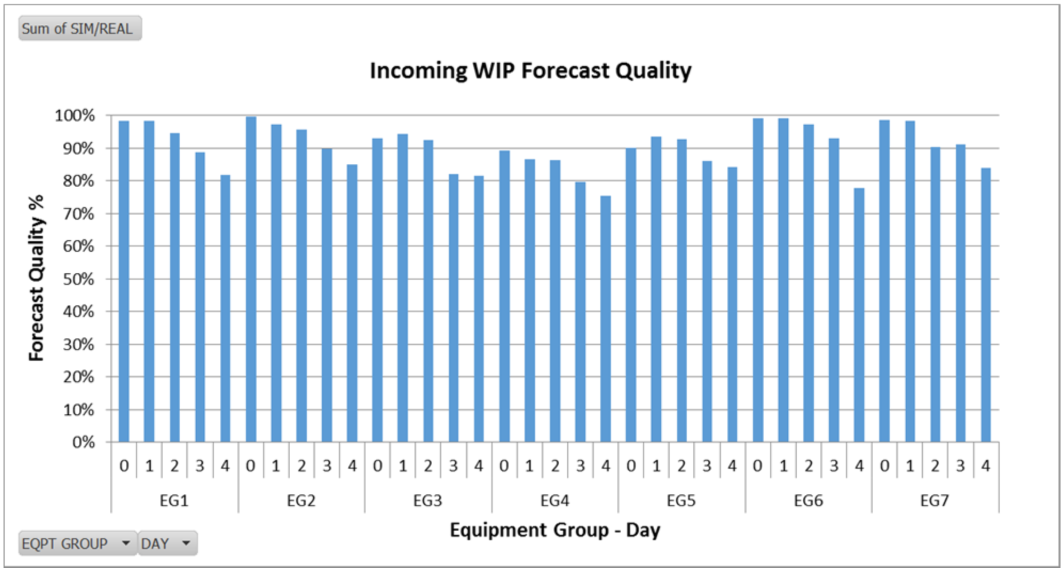


Figure 2: Some equipment groups weekly incoming WIP forecast gap.

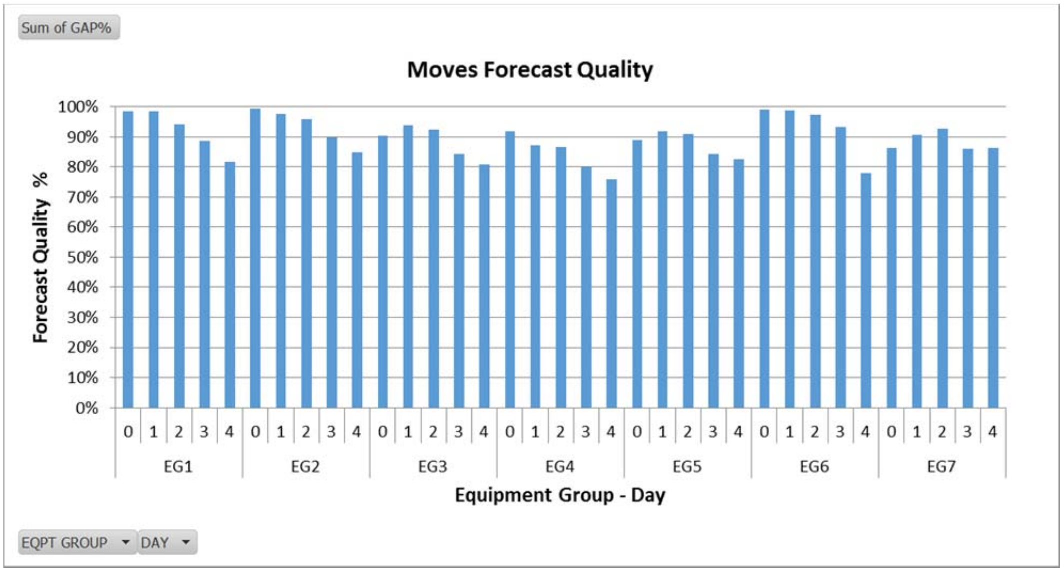


Figure 3: Some equipment groups weekly moves.

4.2 Impact of Process Flow Compression

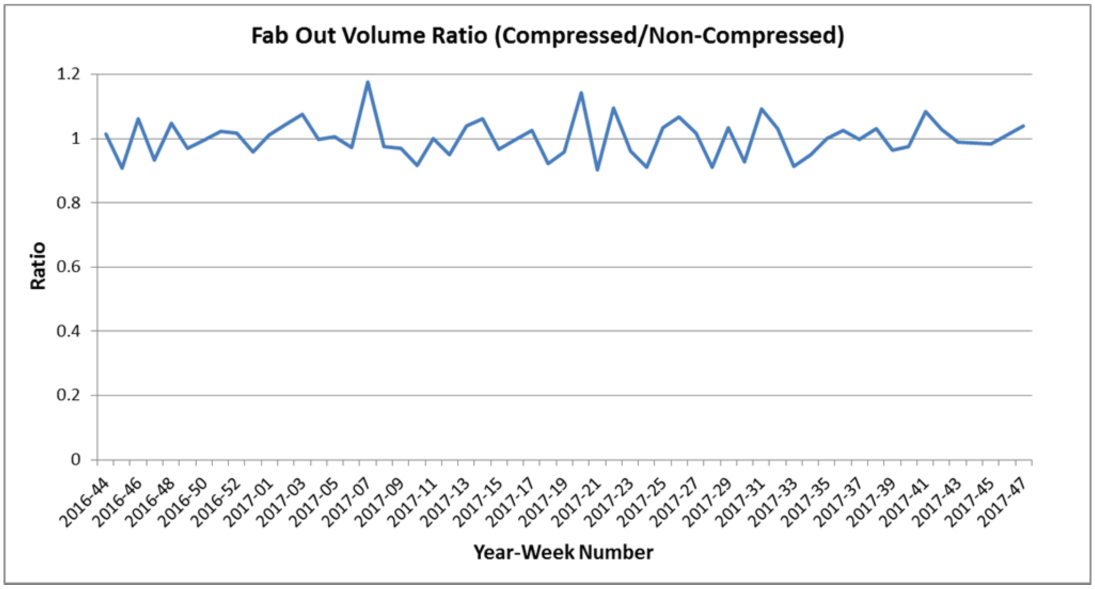


Figure 4: Fab out comparison.

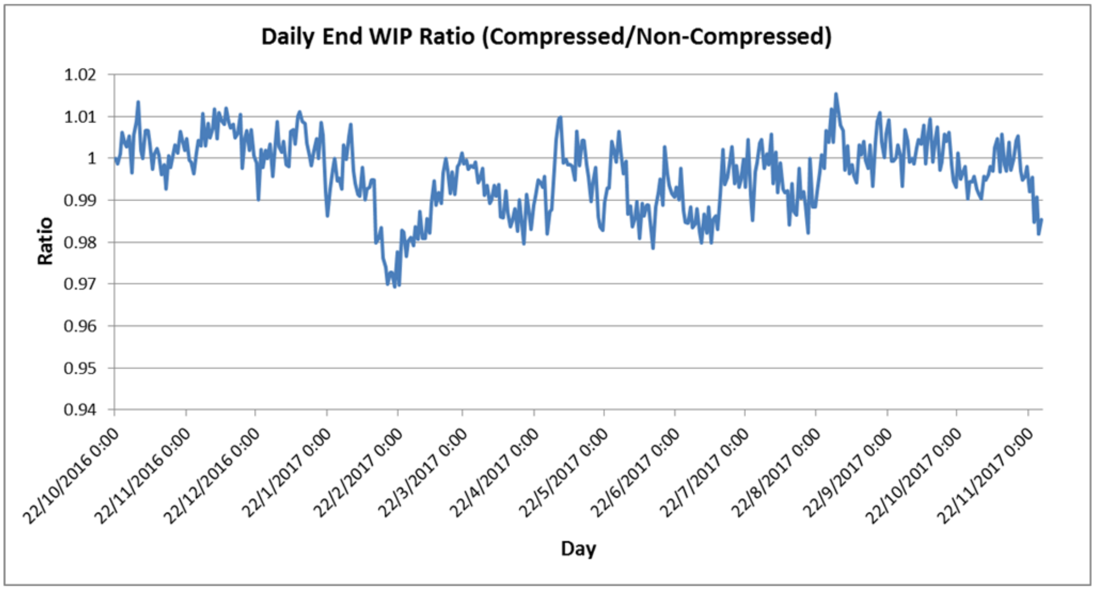


Figure 5: WIP comparison.

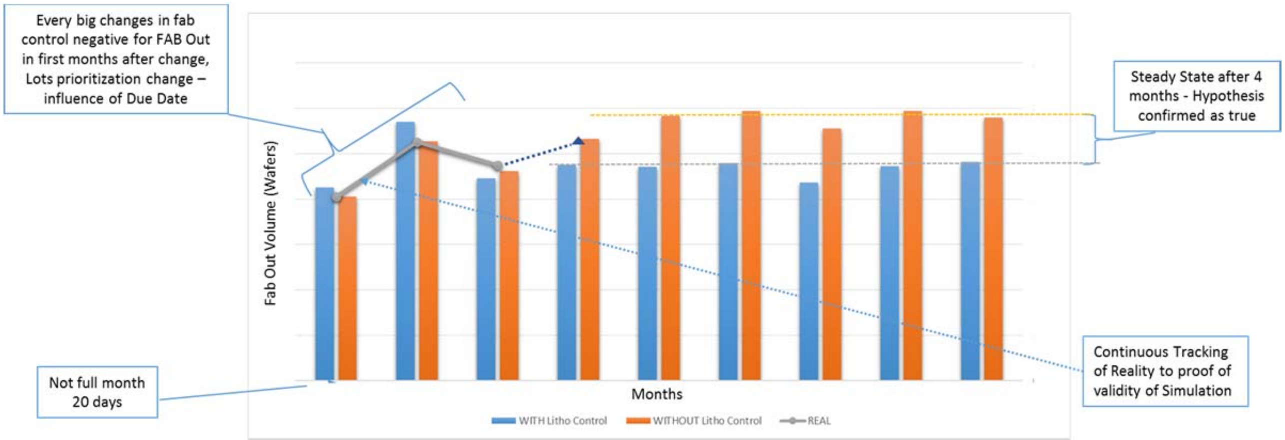
One of the simplifications that was done for long term simulation is process flow compression where some non-critical steps are modeled as unlimited capacity delay steps. Using this approach, it is crucial to ensure that the accuracy of the model is not compromised. Figure 4 and Figure 5 give a comparison of the weekly fab out volume ratio and daily end WIP level ratio between non-compressed and compressed process flows. The ratio is calculated using the formula below:

$$KPI\ Ratio = \frac{Compressed\ Route\ KPI}{Non{-}Compressed\ Route\ \ KPI},$$

where the KPIs are weekly fab outs and daily end WIP level. The two approaches are considered the same if the KPI ratio is 1. The analysis shows that both KPI values stay approximately the same throughout the duration of simulation, with a ratio of between 0.97 to 1.2. The advantage of this method is the significant reduction of simulation run time from 104 minutes to 45 minutes for a 1.5 year run.

4.3 First Scenario Study – Dispatch Rules Evaluation And Realization

One of the first questions the long term model was used to answer was the effect of dispatching when achieving consistency of fab out volumes. Line control observed continuous fluctuations of fab outs over a long period of time. At times, the fab out reached the designated capacity, but other times the fab outs dropped below that level. The line control believed that this could be due to the dispatch rules that control the WIP flow to achieve line balance which in turn creating WIP waves. The long term simulation model was used to evaluate the impact of switching off the WIP flow control dispatch rule. Figure 6 shows the fab out forecast, comparing the scenario with and without a WIP control dispatch rule. The fab out volume dropped in the first month without the WIP flow control dispatch rule. This is because of the sudden increase in WIP flow through the line, resulting in WIP built-up in many areas. Once the WIP wave was cleared, the fab out volume began to improve and the fab simulation consistently observed a higher fab outs volume, comparing with the scenario running with WIP flow control dispatch rule. In that scenario, WIP was building up and thus the lower fab out volume. We have also drawn a line connection at the first 4 bar charts which was an indicator of actual fab out that we observed with the WIP flow control dispatch rule turned off. Today, our fab has consistently reached its design capacity fab outs volume. This contributed to approximately a 20% improvement in fab outs volume. We no longer need to reduce weekly start to manage the WIP built-up due to the WIP control dispatch rule.


Every big changes in fab control negative for FAB Out in first months after change, Lots prioritization change – influence of Due Date

Steady State after 4 months - Hypothesis confirmed as true

Not full month 20 days

Continuous Tracking of Reality to proof of validity of Simulation

Figure 6: Impact of WIP flow control dispatch rule on fab out.

5 CONCLUSIONS

We have successfully extended our short term simulation model to be used for long term simulation studies. Our approach was the simplification of several important modeling elements, which in turn improved the simulation run time significantly. Fast run time is especially critical for long term simulation as a secure run with multiple replications is a must due to high variations contributed by random events. The long term simulation model is now being used extensively to study mid-term wafer start planning, ramp-up scenarios to identify capital investment, and tool qualification releases.

REFERENCES

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

Gan, B.P., P. Lendermann, Y.L. Loh, H.K. Tan, S.K. Lieu, L.F. McGinnis, and J.W. Fowler. 2004. "Analysis of a Borderless Fab Scenario in a Distributed Simulation Testbed." In *Proceedings of the 2004 Winter Simulation Conference*, edited by R.G. Ingalls, M.D. Rossetti, J.S. Smith, and B.A. Peters, 1896-1901. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Narahari, Y., and L. M. Khan. 1997. "Modeling the Effect of Hot Lots in Semiconductor Manufacturing Systems." *IEEE Transactions on Semiconductor Manufacturing* 10(1): 185--188.

Preuss, P., A. Naumann, W. Scholl, B.P. Gan, and P. Lendermann. 2014. "Enhancement of Simulation-Based Semiconductor Manufacturing Forecast Quality Through Hybrid Tool Down Time Modelling." In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S.Y. Diallo, I.O. Ryzhov, L. Yilmaz, S. Buckley, and J.A. Miller. 2444-2453. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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. Yucesan. 2563-2574. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Scholl, W., M. Mosinski, B.P. Gan, P. Lendermann, P. Preuss, and D. Noack. 2012. "A Multi-Stage Discrete Event Simulation Approach for Scheduling of Maintenance Activities in a Semiconductor Manufacturing Line". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher. 2263-2272. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

School, W., M. Förster, P. Preuss, A. Naumann, B.P. Gan, and P. Lendermann. 2016. "Simulation-Enabled Development Lot Journey Smoothening in a Fully-Utilised Semiconductor Manufacturing Line." In *Proceedings of the 2016 Winter Simulation Conference*, edited by T.M.K. Roeder, P.I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S.E. Chick. 2559-2567. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

MARCIN MOSINSKI (PhD) works as a Senior Staff Expert for FAB modeling and simulation at Bosch GmbH in Reutlingen (Germany). He received his M.S. degree in computer science from Dresden University of Technology and his PhD at Munich University of the Federal Armed Forces. His work experience includes Big Data analysis for creation of simulation models by Infineon Technologies and design of new lot dispatch concepts by Globalfoundries. His research interests include simulative and analytical forecasting of complex problems in manufacturing facilities and the statistical data analysis. His email address is marcin.mosinski@de.bosch.com.

TOBIAS WEISSGÄRBER is the Senior Manager for Real Time Dispatching and Line Control of the Robert Bosch Wafer- and Sensorfab in Reutlingen. He has been involved in simulation and production control since three years and was responsible for several fab performance improvement projects. Prior to joining Robert Bosch GmbH, Tobias Weissgaerber was the Group Leader for Production and Lean Management at a Small and Medium Enterprise. He holds a Master of Business Administration from the ESB Business School at Reutlingen University and a Diploma Degree of industrial engineering from the Cooperative State University in Stuttgart. His email address is tobias.weissgaerber@de.bosch.com.

SOO LEEN LOW is a Software Engineer at D-SIMLAB Technologies (Singapore). She is responsible for simulation modelling and analysis of Wafer Fabrication plants. She earned a Bachelor of Engineering in Computer Engineering from National University of Singapore (NUS) in 2014. Her email address is soo.leen@d-simlab.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 a focus on semiconductor manufacturing and aviation spare inventory management. He is now responsible for technology and product development at D-SIMLAB as well as execution of projects in the semiconductor industry, which led to multimillion dollars savings. He holds a Master of Applied Science degree, specializing in Computer Engineering. His email address is boonping@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.