Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

BRIDGING THE GAP BETWEEN FAB SIMULATION AND SUPPLY CHAIN FORECAST

Wolfgang Scholl
Matthias Förster

Infineon Technologies Dresden GmbH
Königsbrücker Strasse 180
Dresden, 01099, GERMANY

Patrick Preuss
André Naumann

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

Hui Ping Ooi
Boon Ping Gan

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

ABSTRACT

The Infineon Dresden 300mm fab is the first wafer fab in the world to run high volume production of power technology products. This requires the coordination of multiple production facilities in the supply chain. Output from one facility is fed as an input to the subsequent facilities in a push or pull operation mode, depending on technology lines. In a push operation mode, lots transition from one facility to another in a continuous manner, while in the pull operation mode, lots are stored and continuation of the lots is triggered by customer order to the supply chain. The latter case triggers the need of synchronizing material flows from one facility to another, and thus requiring an accurate material forecast for better planning and execution. In this paper, we discuss the modelling issues and solutions associated with this challenge, and the use cases of the simulation forecast.

1 INTRODUCTION

Infineon has two highly automated semiconductor manufacturing fabs in Dresden. The 300mm facilities is the first fab in the world that produces high volume power products on 300mm wafers. High complexity in production with a large variations of technologies and products requiring 100 to 1000 production steps to be produced, has to be managed to ensure delivery commitments with short, controllable and predictable cycle times in the manufacturing lines. The 300mm manufacturing site is structured from the view of technology and autonomous manufacturing areas (facilities), e.g. wafer fabrication, backside processing, wafer test and pre-assembly. Depending on the technology line, stocks can be located between the facilities. The material flow in the fab can be discontinued at the interface to subsequent facilities. In our 300mm manufacturing we have two technology lines with different material flows models. In technology line A, material flows from one facility to the next facility without any storage between facilities. This operating mode is thus "push" in nature. On the other hand, the material flow in technology line B is discontinued by a storage between facilities. Lots are started from storage upon customer placing order. This operating mode is thus "pull" in nature. One of the key challenges that we faced was to determine the best time to start lots from the storage to fulfill customer orders, as any started lots will compete for bottleneck machine groups

capacity with lots that flow through in the push operating mode. As such, a performance forecast for the baseload is needed to optimize the lot starts from storage in order to prevent overload situations.

The fab is the first part of the semiconductor supply chain (Ehm et al. 2011). Subsequent supply chain segments need information on anticipated capacity loading as well as the delivery situation in the fab compared to the commitment. Topics as performance monitoring and forecasting cannot be considered insulated only for the fab independently from backend supply chain requests. The backend supply chain needs a full picture from the manufacturing site. In wafer fabrication facility of the 300mm fab in Dresden discrete-event simulation is a widely used method in daily business for KPI forecasting as well as scheduling of activities in operations. Former fab simulation activities were more or less focused on wafer fabrication facility only. Subsequent facilities (backside processing, wafer test, preassembly) are not considered yet. That means, there is a gap between dynamic KPI forecast of the wafer fab and the interface to backend supply chain outside of the Dresden fab. In this way, the benefits of fab simulation methods for dynamic capacity planning cannot be applied to meet the requirements from the perspective of the supply chain. Innovative planning and forecasting methods, firstly developed and used for wafer fabrication facility, have to be rolled-out to all facilities of the manufacturing site.

2 MODELLING APPROACH

On its way through the whole production cycle each lot in the 300mm line goes through different facilities. Each lot belongs to one out of two major technology lines. Each of them results in a different facility transition sequence. Figure 1 shows the basic structure of those technology lines and some specialties which needs to be considered during simulation model building.

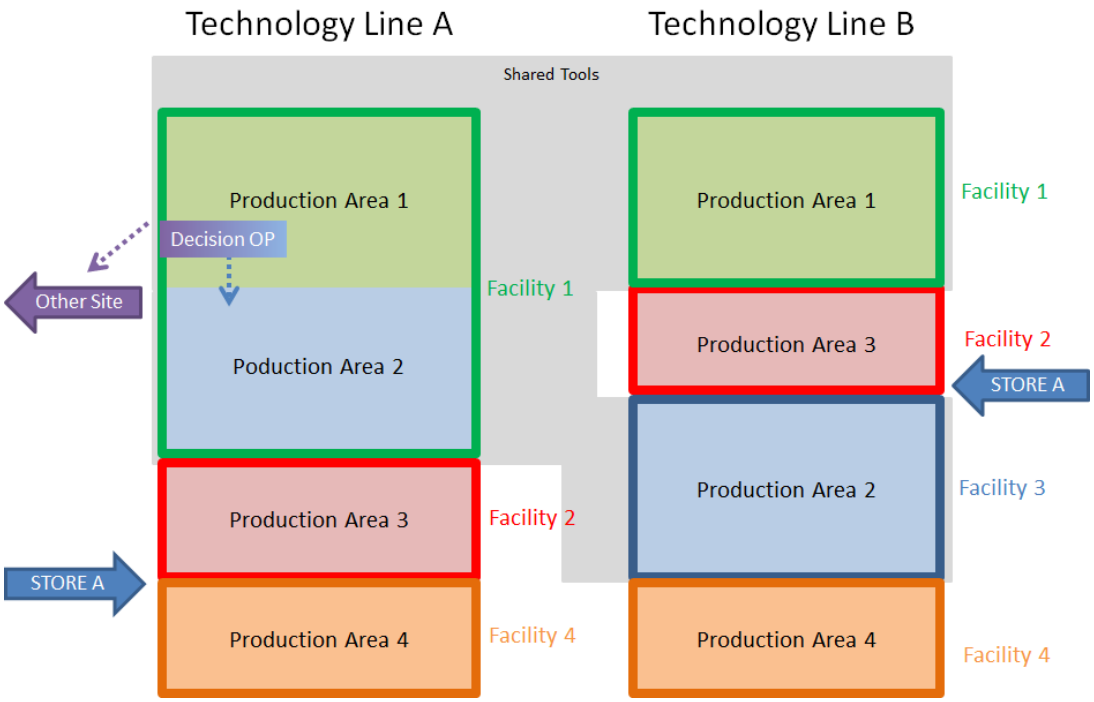

Figure 1: Facility transition modelling for both Technology Lines.

2.1 Push-Pull Principle

One aspect is the contradicting principles in the operating mode of material flow. Referring to Figure 1, technology line A is a continuous material flow over all facilities (push mode), the material flow of

technology line B is discontinued by storage between facilities. Lots are started from storage based on customer orders and pulled to transition into the next facility of the supply chain. One of the pressing issue in the pull operating mode is related to situations where demand is higher than material in storage. If this situation is not anticipated, there will be a risk of missing customer delivery deadlines as it would be too late to rush the facility before to supply the lots quickly. To mitigate this risk, a simulation based forecast is required, and will be discussed in details in Section 4.2.

2.2 Transition Modelling

In our simulation model the pull driven facility transitions are modelled by adding additional process steps at the end of the previous model and let the lot continues its journey on the next route. Transitions which are driven by the pull principle are modelled using lot start events based on available wafer start plans. In addition, due to the fact that each facility has its own set of unique process flows and each lot can visit more than one facility on the way through the extended facility model, we needed to support the possibilities of modelling a chain of process flows. This chain defines the relationship of process flows between facilities.

2.3 Modeling Fidelity

To decide how the fidelity of new facility models needs to be chosen, a wide set of real data assessment has been conducted, covering data availability analysis, data quality assessment and lot cycle time analysis to support the final modelling decision for each of the new facilities.

2.3.1 Data Availability Analysis

Initial step on the data assessment activities was to check if all the required data was available for the new facilities. The result has been consolidated as shown in Figure 2, where the green cell indicated that the data is available, while cell that marks with "Derived from History" means that the data has to be generated from historical data analysis.

Facility	Lot Trace	Equipment State Trace	Route	Equipment	Timing	Dedication	Lot Starts
Facility 1							
Facility 2						Derived from History	Derived from History
Facility 3						Derived from History	Derived from History

Figure 2: Summary of available simulation input data for the new facilities.

Dedication defines processes that are able to run on specific tools. Deriving this information from historical data is straightforward as we have data available that shows which process ran on which tool. The only shortcoming of this approach is that any new processes would not have run before, and thus there will be no dedication at all. Our solution to this issue is to map similar processes to the same tool or tool group. This will at least ensure that there are run paths for all started lots.

Due to the nature of the technology line B, the availability of lot start plan is limited as customer places order in short notice, and expecting the order to be fulfilled in a short time. The lot start plan changes by the day, and imposes great challenges on the modelling. To address this issue, we decided to derive the lot

start plan from historical data, which basically means we are generating volume and product mix profile similar to the past two weeks. Our reference does not go too far into the past because the product mix profile would have changed quite significantly.

2.3.2 Cycle Time Analysis

To model the process flow within a facility, we have two options. The first option is to model each facility process flows in details, while the second option is to model the whole facility with a cycle time distribution derived from historical data. Referring back to Figure 1, we had to model the process flow of *Production Area 1* and *Production Area 2* in details. *Production Area 1* and *2* are basically the wafer fabrication facility which share tool capacity. It would not make sense to model them as cycle time distribution as we do have sufficient data to build the simulation model to achieve good forecast accuracy. For one of the other facilities, we faced the issue of obtaining data associated with product capacity needs and thus had no choice to model the facility as cycle time distribution. Figure 3 shows an example of a process time distribution for one of the facilities. This distribution is fed to the simulation model without modelling each tool or resources with capacity limit. With this modelling approach, we are also running simulation with replications to ensure good confidence interval of the observed results.

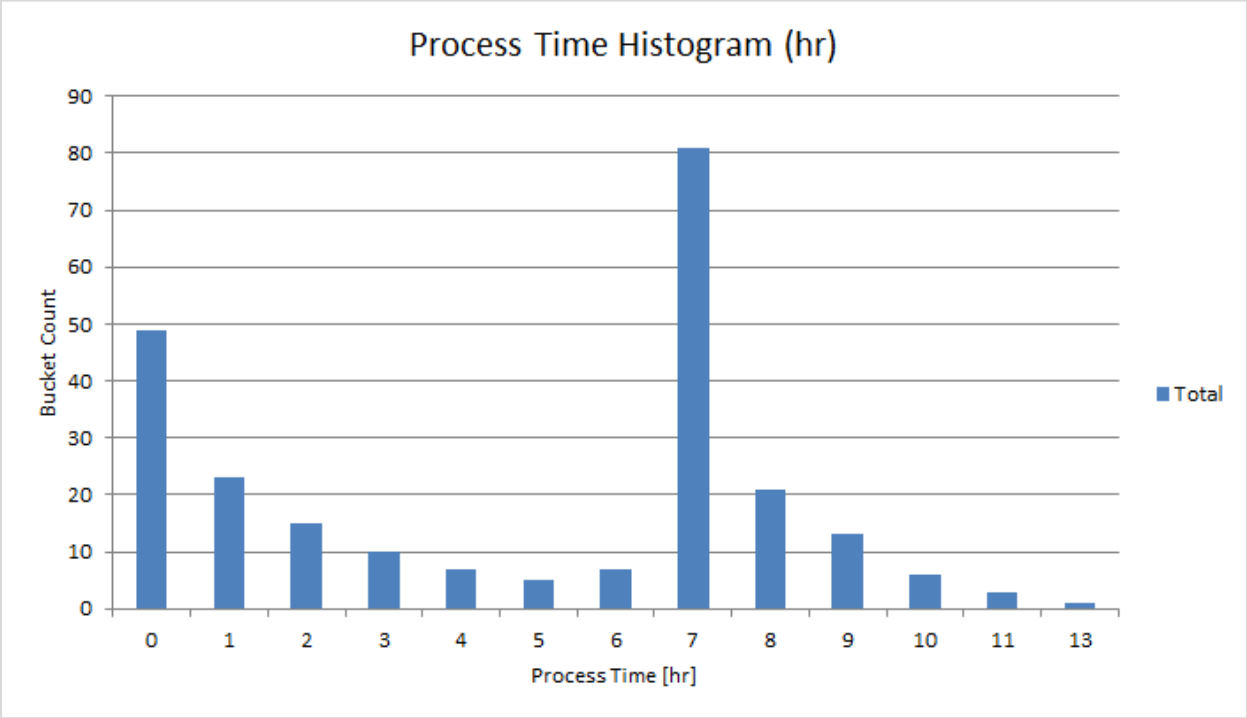

Figure 3: Facility process time histogram with high variance resulting in delay step modelling.

3 CHALLENGES

For the process of model building several data issues need to be addressed. A proper handling is crucial for high quality forecast.

3.1 Tool Sharing

There are tools which are dedicated to one facility, while others shared between facilities. Extending the model to include all facilities had a positive impact on the forecast quality of our existing fab simulation

model. This also implies that the overall model forecast quality is highly influence by the right modeling abstraction of each facility. Bad forecast quality in one facility will have negative impact on the forecast quality of another facility. This makes the process of validation very complex as model calibration to address a modelling shortcoming of one facility might have a negative impact on another.

3.2 Assessment Lots

Lots travelling within the new production areas sometimes need to be stopped for investigation by the process engineer ("Assessment Lots"). This delay is not part of the standard process time measured on a processing step. This is important to be considered as part of the simulation model building to achieve a correct delivery forecast for subsequent facilities. Figure 4 illustrates the waiting time distribution for those assessment activities derived from historical data of the last three months.

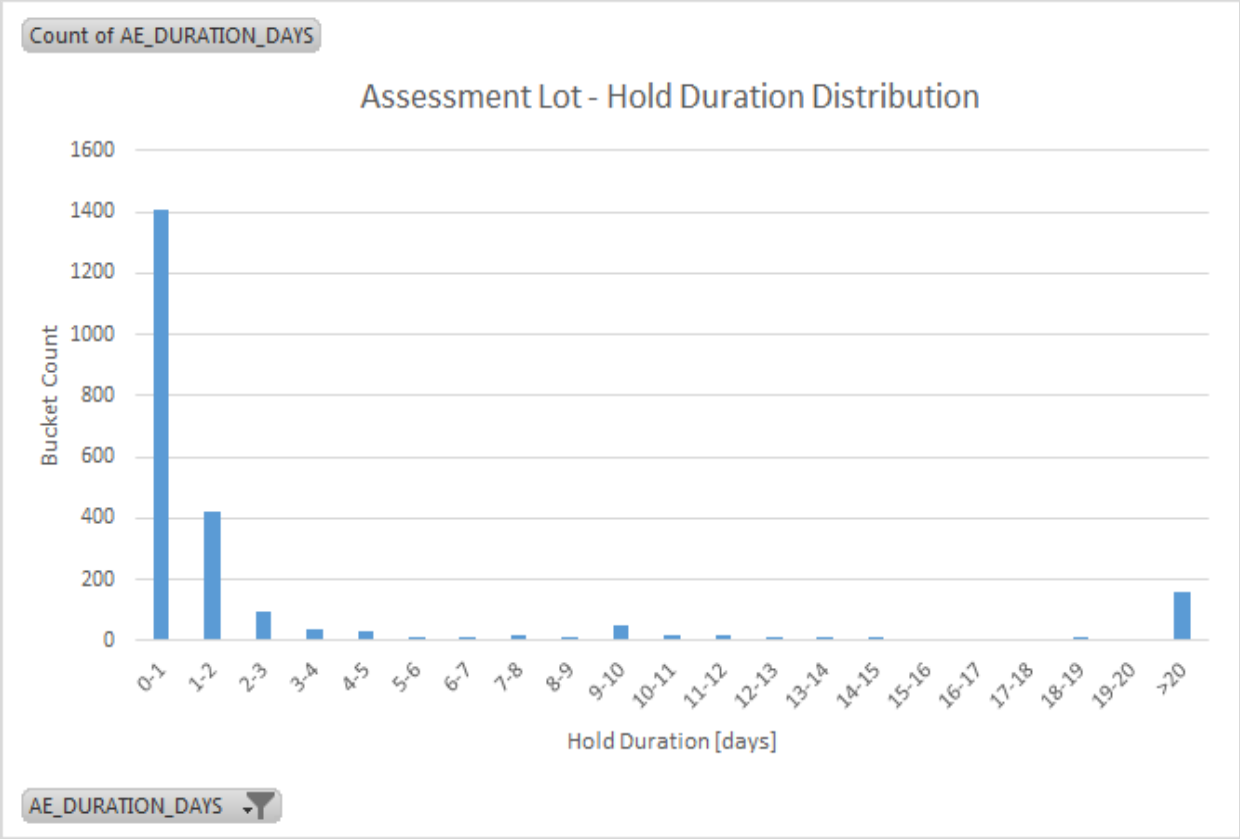

Figure 4: Lot assessment waiting time distribution.

This time distribution was added to the model to mimic the lot waiting times for assessment. However, the challenge comes when we need to forecast such events happening in the simulation future not known at the time of work-in-progress initialization. There is no well-defined condition on when and for which lots such assessments will be required, thus a condition based modelling element is not possible. In order to reflect the correct number of moving lots in the system an alternative approach is applied. Figure 5 shows the typical daily number of lots in a hold state due to ongoing experiments. As seen in the chart, variation of this number is low. Thus, we approximated this behavior by ensuring that the number of lots in hold state stays at approximately the average observed value. This helps to ensure the average performance forecast of the model is accurate. But the shortcoming of this approach is that any lot level forecast is not feasible as we might have held back a different lots in simulation and reality.

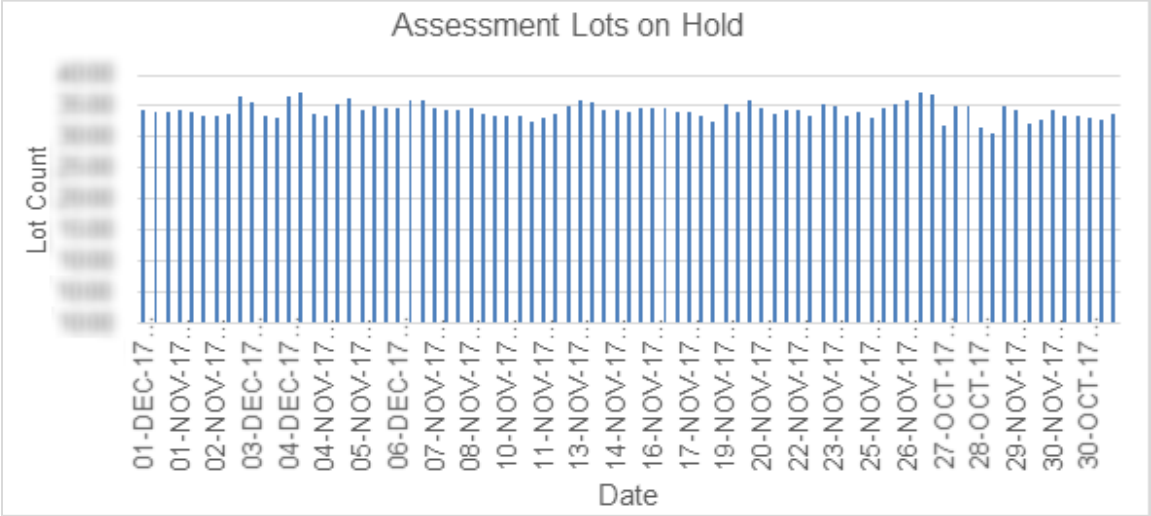

Figure 5: Daily number of lots on hold state due to ongoing analysis by the process engineer.

3.3 Demand-Based Wafer Start Plan

As explained in Section 2, some of the facility is driven by a lot start plan. The lot start plan must be accurate to achieve good forecast quality. The left set of charts in Figure 6 shows an equipment group forecast with incomplete lot start data. This equipment group is near to the beginning of the facility process flow. As observed, the three key KPIs, wafers in, wafers out, and WIP, were not forecasted well. There was major gap between the reality (blue bar) and simulation (red bar) values. The right set of charts in Figure 6 was driven with a corrected lot start plan. It is obvious in this illustration that having the right lot start plan is essential in achieving good forecast quality.

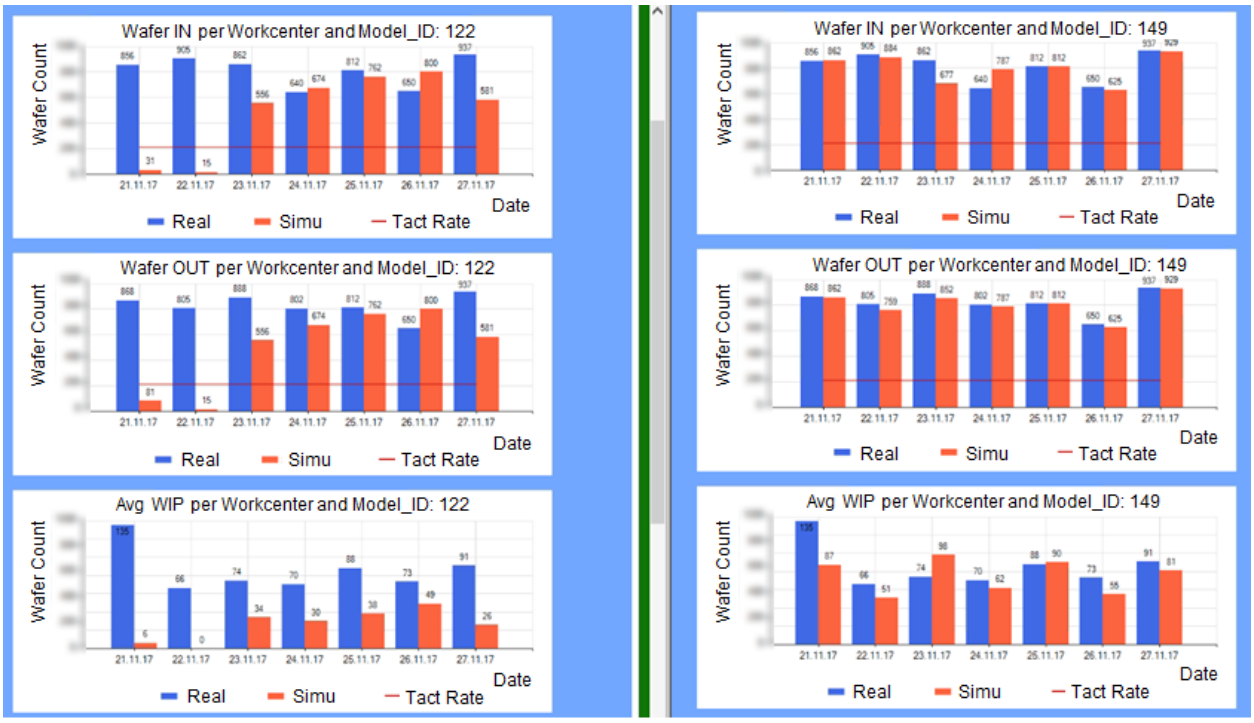

Figure 6: Equipment Group forecast quality with incomplete (left) and corrected wafer start plan (right).

4 SPECIAL USE CASES

Several new use cases have been established as part of the facility extension activities. The focus in this paper is the forecast of lot flow leaving or starting at a new facility.

4.1 Facility based KPI forecast

Extending the standard work center KPI reports (Scholl 2011 and Seidel 2017) of the wafer fab facility model, we added another layer of data to the forecast reports. Figure 7 illustrates the new report design, with each bar chart of KPIs such as arrival, move, or WIP, stacks with forecast for each facility. This enables the user to know the mix of capacity consumption for each facility, and enabling actions when there is an unexpected high capacity consumption by one facility that might jeopardize the performance of another. The ultimate objective is to ensure production for all facilities can proceed smoothly without running into unnecessary bottleneck situations, as recovering is going to have negative impact on overall fab performance.

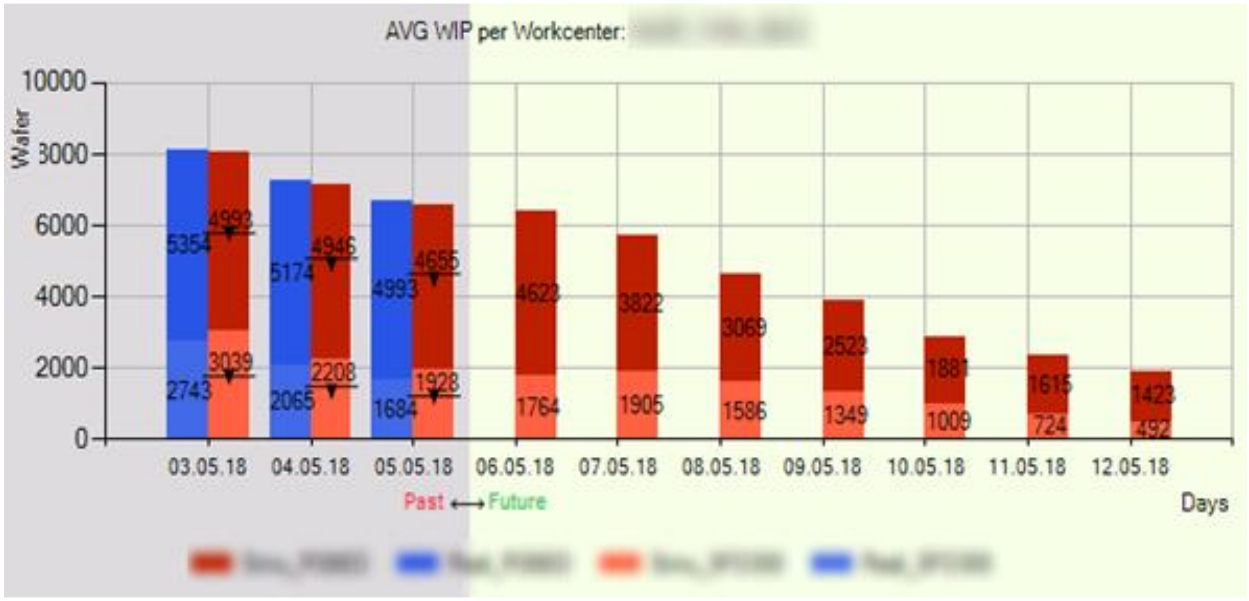

Figure 7: Work Center WIP forecast with additional facility information in the stacked bar chart.

4.2 Early warning for over/under-utilized capacity at facility transition storage

As discussed earlier, the technological line B is produced with storage in between subsequent facilities. Lots are started from storage triggered by customer orders and being pulled to transition through the supply chain. This triggers the needs of a baseload forecast, to provide a guideline in optimizing lot starts from the storage to prevent overload and material shortage situations. Figure 8 gives an illustration of a cumulative storage report. The simulation forecast provides the material output forecast and thus the material volume incoming to the storage. The lot start plan will thus consume this material to fulfill customer orders. If net effect is positive, we do not have a material shortage situation. Otherwise, we run into a material shortage situation, and the previous facility will be triggered to replenish the material faster. Using this approach, we were able to reduce the situations of running low in material for lot start in the next facility, and any shortage is triggered before the situation actually arises. This more proactive approach of managing the material level eliminates a lot of fire-fighting situations that could have arose. Ultimately, the accuracy of this forecast relies heavily on the accuracy of the lot start plan, and the material out forecast of the simulation.

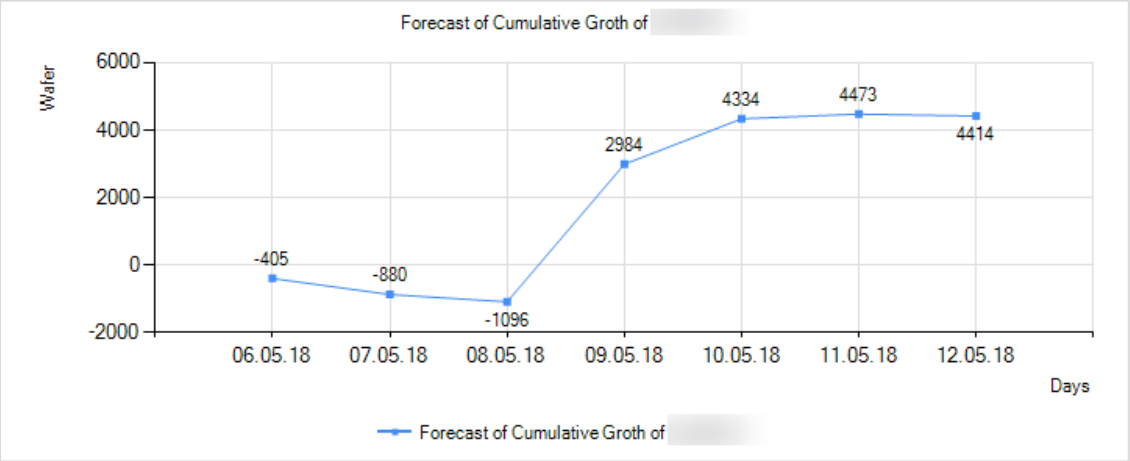

Figure 8: Cumulated storage inventory considering incoming stream derived from simulation results and outgoing stream taken from wafer start plan.

5 CONCLUSION

Our short-term simulation model has been successfully extended to include manufacturing facilities beyond the wafer fabrication plant, bringing our model closer to a supply chain simulation model. The primary challenge that we faced during the process of validation has been the interlocking forecast quality between facilities due to sharing of resource capacity, and the accuracy of lot start plan that is highly influenced by unpredictable customer demand. We took the approach of deriving future lot start plan from historical data, making assumptions of small mix changes over several weeks. The model has been able to provide good guidelines to the production to avoid out-of-material situation to start production in subsequent facilities. This also smoothen the WIP flow situation as the forecast enables reduction of fire-fighting situation, that typically involve quickly pushing WIP lots through that could create WIP bubble which is not healthy for line performance.

REFERENCES

D-SIMLAB Technologies. 2018. Forecaster and Scenario Manager. http://www.d-simlab.com/category/d-simcon/products-d-simcon/forecaster-and-scenario-manager, accessed August 24, 2018.

Ehm, H., T. Ponsignon, and T. Kaufmann. 2011. "The Global Supply Chain Is Our New Fab: Integration and Automation Challenges". In *Proceedings of the 2011 Advanced Semiconductor Manufacturing Conference*, 1-6, New Jersey: IEEE, May 16th-18th, Saratoga Springs, NY, USA.

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*, edited by S. Janet al. 1932-1943. Piscataway, New Jersey: IEEE.

Seidel, G., B. P. Gan, C. W. Chan, C. F. Lee, A. M. Kam, A. Naumann, and P. Preuss. 2017. "Harmonizing Operations Management of Key Stakeholders in Wafer Fab Using Discrete Event Simulation". In *Proceedings of the 2017 Winter Simulation Conference,* edited by W. K. V. Chan et al. Piscataway, New Jersey: IEEE.

AUTHOR BIOGRAPHIES

WOLFGANG SCHOLL works as a Senior Staff Expert for modeling and simulation for Infineon Technologies in Dresden (Germany). He 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 in the field of capacity planning. Since 2003 he is responsible for fab simulation. He supervises development and roll-out projects and is also a member of the Supply Chain Simulation community. His email address is wolfgang.scholl@infineon.com.

MATTHIAS FÖRSTER works as a System Expert for modeling and simulation for Infineon Technologies in Dresden (Germany). He studied micro technology at the University of Applied Sciences Zwickau (Germany) and graduated with a diploma degree. Upon joining Infineon Technologies in Dresden in 1996, he first worked in the Plasma Processes Department. In 2001 he changed to the Furnace Processes Department as a System Expert. Since 2014 he works in the simulation team and is responsible for short term forecasting. His email address is matthias.foerster.drs@infineon.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, Infineon and Bosch 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 Senior 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 at D-SIMLAB Technologies (Germany) in deployment projects for customers in semiconductor industry. His email address is andre@d-simlab.com.

HUI PING OOI is Data Analyst at D-SIMLAB Technologies (Singapore). She joined D-SIMLAB in 2015 with focus on simulation modeling and validation within semiconductor manufacturing. Irene graduated from the National University of Singapore with a degree in Industrial Engineering. Her email address is irene.ooi@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 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 holds a Master of Applied Science degree, specializing in Computer Engineering. His email address is boonping@d-simlab.com.