Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

SIMULATING FRONT-END SEMICONDUCTOR SUPPLY CHAINS TO ASSESS MASTER PLANS UNDER UNCERTAINTY: A CASE STUDY

Aaron Sieders[1], Cas Rosman[2], Collin Drent[3], and Alp Akcay[4]

[1] Sales Operations, NXP Semiconductors N.V., Eindhoven, NETHERLANDS
[2] Dept. of Supply Chain Innovation, NXP Semiconductors N.V., Eindhoven, NETHERLANDS
[3] Dept. of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, NETHERLANDS
[4] Dept. of Mechanical and Industrial Engineering, Northeastern University, Boston, USA

ABSTRACT

This research presents an aggregated simulation model for the front-end semiconductor supply chain to assess master plans, focusing on the impact of demand and supply uncertainties on the key performance indicators on-time delivery and inventory on hand. Supply uncertainty is modeled using discrete distributions of historical cycle times, incorporating load-dependent cycle times through a non-linear regression model. To model demand uncertainty, we use future forecasts and adjust them by sampling from distributions of historical forecast percentage errors. By comparing master plan performance under uncertain conditions with those from deterministic scenarios, the model provides valuable insights into how these uncertainties influence supply chain performance. Using data from NXP Semiconductors N.V., a Dutch semiconductor manufacturing and design company, we demonstrate the model's applicability and offer practical guidance for industry practitioners. Based on numerical experiments, we conclude that the impact of demand and supply uncertainty significantly differs compared to deterministic planning.

1 INTRODUCTION

Semiconductors, the essential components for all electronic devices power everything, ranging from smartphones and computers to advanced medical equipment and cutting-edge artificial intelligence systems. They have driven advances across various technologies and contributed to economic growth and global competitiveness. The semiconductor industry is one of the fastest-growing industries, with semiconductor components being essential, albeit in varying degrees, to all technology products. The semiconductor supply chain comprises a front-end and back-end process. The front-end processes involve fabricating microchips from a blank wafer to a completed wafer, where the microchips are created but remain on the wafer. The back-end process involves assembly, where the semiconductor is separated from the wafer and transformed into the end product. The die bank is an inventory point between the front-end and back-end processes, where semiconductor wafers are stored before being assembled into final products.

The semiconductor industry faces complex master planning and control challenges (Mönch et al. 2013), which have grown increasingly significant for semiconductor supply chains over time (Chien et al. 2008), (Mönch et al. 2018). Master planning adopts a firm-wide, medium-term perspective on determining what to produce, where, and how. Due to the scale and complexity, generalizations and assumptions, such as treating demand forecasts, production capacities, and cycle times as deterministic, are necessary. The resulting deterministic master planning problem is then often solved with mathematical programming models, most commonly linear programs, due to their ability to optimize complex decision-making processes involving multiple variables and constraints (Leachman 2002).

Cycle times, the time between work being released into the fab and its completion as a finished product, are thus assumed to be deterministic and independent of resource utilization. However, queuing theory

979-8-3315-8726-0/25/$31.00 ©2025 IEEE

(Curry and Feldman 2011), simulation models (Atherton and Atherton 1995), and industrial observation (Wu 2005) all indicate that the mean and variance of cycle times increase non-linearly with resource utilization. Furthermore, forecasting demand is challenging due to volatile market demand and short product life cycles (Wang and Chen 2019). Deterministic mathematical programming models thus facilitate the incorporation of complex technological constraints but ignore the pervasive stochastic aspects of both the production process and demand (Ziarnetzky et al. 2020). This lack of consideration for uncertainty can result in plans that are not robust to unexpected events. The robustness of a master plan refers to its ability to remain effective and achieve its objectives when faced with uncertainty in demand and supply. Moreover, taking uncertainty into account can yield substantial variations in performance indicators compared to deterministic planning, highlighting the limitations of solely relying on deterministic planning results.

In this study, we propose an aggregated simulation model of the front-end manufacturing process to assess master plans and deviations in the expected performance due to uncertainty. The proposed simulation model is based on the existing master data structure of the company, limiting data maintenance effort and time to set up the simulation model compared to a more detailed simulation approach. Additionally, the proposed simulation model only models key bottleneck resources, reducing the computational expense. Moreover, it accounts for load-dependent cycle times. We demonstrate the use of the aggregated simulation model by evaluating master plans under stochastic supply and demand conditions and comparing the performance with the expected deterministic outcomes. The main contribution of this paper is a methodology for simulating front-end semiconductor supply chains using a high-level, aggregated modeling approach that also includes load-dependent cycle times.

The paper is organized as follows: first, relevant literature in Section 2 is discussed, then we propose our method in Section 3. Next, we apply the method to a real-world use case in Section 4, followed by a discussion in Section 5, the managerial insights in Section 6, and finally, the conclusion in Section 7.

2 LITERATURE REVIEW

Extensive studies have been conducted on the simulation of semiconductor supply chains. Based on our review of simulation models for semiconductor supply chains, we identify three main streams for this purpose: (1) using system dynamics (SD), (2) using reduced simulation models within a detailed simulation model, and (3) using only historical data.

(1) System dynamics. Orcun and Uzsoy (2011) use SD to study the effect of Production Planning on the dynamic behaviour of a simple semiconductor supply chain. Hartwick et al. (2023) assess the effects of external disruptions on a simplified semiconductor supply chain, and evaluate mitigation strategies. System Dynamics models allow for a reduced computing effort and effectively capture feedback loops, time delays, and nonlinear behaviors in complex systems like semiconductor supply chains.

(2) Reduced simulation models. Detailed simulation models capture the intricate behavior and dynamic processes of a wafer fabrication (wafer fab) facility. Kopp et al. (2020) introduces four detailed discrete-event simulation models representing modern wafer fabs. The aim of the testbed consists in providing researchers with a platform able to credibly represent the complexity of modern semiconductor manufacturing. However, detailed simulation models require large amounts of data and long computation times to produce statistically valid results (Fowler et al. 2015). Therefore, reduction techniques have been introduced to mitigate these challenges, such as modeling only bottleneck work centers (Hung and Chang 1999). Rose further studies a similar approach in a series of papers (Rose 1999), (Rose 2007a), and (Rose 2007b). To make the model utilization dependent, they replace the fixed delay time distributions in the delay units with delay time distributions depending on the current inventory level. Ewen et al. (2017) use a similar bottleneck-based simulation approach. However, in contrast to other work, they linearly interpolate between different load situations to account for load-dependent cycle times. Furthermore, they use a second reduction approach inspired by Duarte et al. (2007) that considers load-dependent cycle times. The disadvantage of model reduction approaches is their dependence on detailed simulations to create response surfaces and identify bottleneck workstations. Since developing and maintaining a detailed simulation to obtain the reduced model

requires considerable resources in practice (Shanthikumar et al. 2007), we argue that these approaches are inappropriate.

(3) Using only historical data. The Effective Processing Time (EPT) modeling concept by Spearman (2014) can also be used to simulate the front-end processes. Deenen et al. (2024) developed an aggregated wafer fab model that uses this EPT concept. Although this study has promising results and does not require a detailed simulation, the aggregated model cannot accurately predict the cycle time distribution for the complete wafer fab. Morrice et al. (2005) describe a DES model developed for Freescale Semiconductors using probability distributions for the cycle time estimated from historical data to avoid modeling the supply chain nodes in detail. A limitation of their approach is the assumption that the cycle time distribution is independent of resource utilization or that the utilization level will remain relatively constant over time. In this study, we relax this assumption by including load-dependent cycle times in the simulation model.

In conclusion, research on simulating semiconductor supply chains has been of great interest in the literature. Many studies focus on reduced simulation models to lower computational demands while preserving accuracy. However, these methods often require detailed models to generate data, making them less practical. Therefore, the third category is the most relevant to this study as the approach does not initially require building a detailed simulation model. In this study, we use discrete event simulation (DES) because it enables us to model the stochastic, event-driven behavior of front-end semiconductor manufacturing systems with a level of detail that system dynamics (SD) cannot achieve. The techniques in the second category for capturing load-dependent cycle times are valuable for this study, which relies on historical data rather than detailed simulation models. The approach of Morrice et al. (2005) is highly relevant to this study, as it captures stochasticity within the production processes at a higher aggregation level without needing to model each individual process step. Furthermore, this approach requires minimum effort to set up, which makes the approach practical and scalable. The approach of this study will provide a more realistic representation of the system's performance under varying conditions by including load-dependent cycle times.

3 METHODS

This section describes the method used to model front-end processes and the method used to generate the final demand. First, we develop an aggregated simulation model of the front-end processes to simulate the supply of wafers to the die bank. Second, we model supply uncertainty by leveraging historical cycle times and Work-In-Progress (WIP) levels. Finally, our approach involves modeling demand uncertainty using future forecasts and distributions of historical percentage errors.

3.1 Simulation Setup

To simulate the front-end processes, we use discrete event simulation (DES). Figure 1 depicts an example of the front-end supply chain consisting of wafer fabrication and wafer test. Between these stages, products are transported to the next stock point.

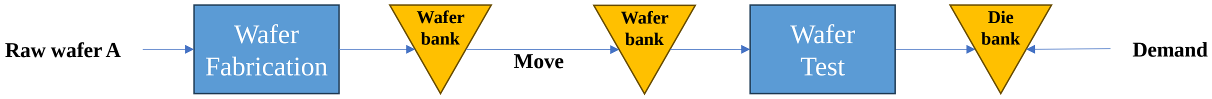


Figure 1: Front-end supply chain model.

Table 1 shows the relevant input parameters for the simulation model. Both wafer fabrication and wafer testing require the same parameters. Each product is associated with a resource in the two processes, each with its cycle time and yield. Only the key bottleneck resources are modeled, one for each process step. The capacities for all resources are specified every week. These parameters and information for each product, including their corresponding resources, can be found in the company's existing master data. The transportation time between the two wafer banks can also be found here. This transportation time is assumed to be fixed over time. This assumption is reasonable given that these times are guaranteed and minor deviations have minimal impact on supply chain performance. We assume that the capacity is fixed over time, as the facility must ensure it meets the stated capacity. This implies that we do not consider uncertainty, such as breakdowns or fluctuations, in capacity availability. Although the front-end supply chain includes multiple wafer fabrication and testing facilities, the operational sequence for products manufactured at different locations remains consistent. The simulation model operates at a daily granularity, as the cycle times are provided in days. A simulation run models the supply execution of a master plan over multiple weeks while incorporating the arrival of demand.

Table 1: Supply chain parameters.

Process step	Input parameters
Wafer fabrication Wafer test	Location
	ResourceID
	Weekly capacity
	Cycle time
	Yield
Move	Transportation time

3.2 Modeling Supply Uncertainty

Similar to the approach of Morrice et al. (2005), this study models supply uncertainty using probability distributions for cycle times, estimated from historical data. This method avoids the need for detailed modeling of each supply chain node. To effectively model supply uncertainty, we first analyze the cycle times of wafer fabrication processes to determine if they exhibit load-dependent behavior. When load-dependent cycle times are identified, we propose using non-linear regression models based on Work-In-Progress (WIP) levels. Given the aggregation level of the simulation model, non-linear regression models offer a practical balance between simplicity and effectiveness for predicting future cycle times, despite not being the most advanced method available. If load dependency is not observed, we fit discrete distributions to cycle times. Load-dependent cycle time modeling is applied exclusively to wafer fabrication and not to wafer test.

In wafer fabrication, each wafer is categorized into a specific wafer technology group based on its technological and processing requirements. The cycle time distributions are analyzed at this group level. We have gathered historical data on cycle times for each lot within the wafer technology group and data on the total WIP levels in the wafer fabrication facilities. Data points outside the 1.5 times interquartile range are considered outliers and removed. This outlier removal technique is also employed when fitting other cycle time distributions. We use a polynomial regression model to understand the relationship between WIP levels and the expected cycle time and 5-fold cross-validation is performed to determine the optimal degree. Polynomial models are chosen for their flexibility in capturing smooth, nonlinear trends commonly observed in manufacturing systems, without requiring a predefined functional form.

To model supply uncertainty, we sample from the discrete residuals of the polynomial regression model and added to the predicted cycle time of the regression model. The residuals are calculated by taking the difference between the actual data points and the estimated value of the polynomial regression model. The residuals are divided into equally broad WIP buckets, as increased WIP levels may lead to greater supply uncertainty. This method introduces variability into the predicted cycle time values, creating a distribution encompassing both the load-dependent cycle time predicted by the regression model and the variability captured by the residuals.

In cases where there is no relationship between WIP levels and cycle time, indicating stable cycle times across different load levels, we adopt an alternative approach to model cycle time variability. Specifically, an empirical distribution is fitted to the differences between planned cycle time (cycle time stated in master data) and actual cycle time. This deviation is then sampled and added to the stated cycle time to determine the cycle time of a lot. Furthermore, we fit an empirical distribution to the yield of the wafer fabrication process. Finally, empirical distributions are fitted to both cycle times and yields of the wafer test production step.

3.3 Modeling Demand Uncertainty

Ponsignon and Mönch (2014) generate final demands D_{pt} for product p in period t by adding a forecast error R to the forecast F_{pt} of the product for in that period, see Expression (1). They assume that this error is normally distributed. In this research, we also employ this method to model demand uncertainty since we are interested in the impact of forecast accuracy. Following the methodology of Rosman et al. (2024), we determine the distribution of the error term using historical data of forecasting percentage errors. Given the monthly granularity of the forecast data, the t denotes a monthly period. It is worth noting that the future forecasts remain unchanged during the simulation run, so a rolling horizon approach is not employed. In the simulation, the final demand for a month is determined by adding a randomly sampled error to the future forecast for that month.

$$D_{pt} = F_{pt} \cdot (1 + R) \tag{1}$$

To determine the distribution of this error term, we first analyze the historical error percentages. We adopt the same methodology as Rosman et al. (2024) to calculate these error percentages. Given that their study focuses on end-product demand, while ours examines the front-end supply chain, we begin by aggregating end-customer demand to wafer-level demand. The forecast percentage errors are calculated for each wafer for each look ahead period. A positive forecast error indicates under-forecasting, where actual demand surpasses estimates. Whereas, a negative forecast error suggests over-forecasting, meaning the actual demand was lower than expected. For each look-ahead period and wafer combination, the historical percentage errors are fitted to five candidate distributions: Normal, Lognormal, Gamma, Weibull, and Beta. The best fit is selected based on the Kolmogorov-Smirnov test, using the highest p-value above the significance threshold of 0.05.

In each simulation run, error percentages are sampled from these distributions and are added to future forecasts to generate final demands. Given that the demand has a monthly granularity while the simulation operates daily, the final monthly demand is evenly distributed across the weeks. Since we do not model the back-end operations, we assume a 'perfect' back-end with sufficient capacity to process the requested demand. This assumption is reasonable, as in front-end master planning, the back-end capacity constraints are often relaxed to ensure that the back-end does not become the bottleneck.

4 APPLICATION

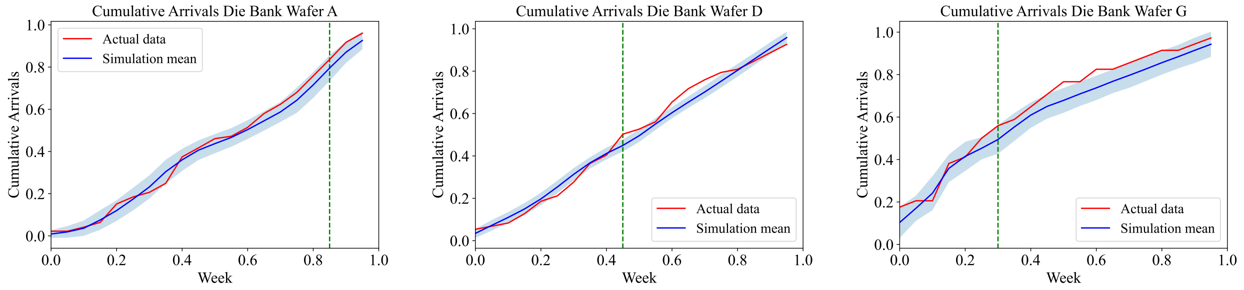
In this section, we apply our method to evaluate the impact of supply and demand uncertainty on key performance indicators, specifically On-Time Delivery (OTD) and Inventory on Hand (IOH), in the die bank. This evaluation aims to emphasize the differences between deterministic planning and planning that accounts for uncertainty. The proposed method is applied to the NXP front-end supply chain. First, the use case is discussed. Subsequently, we examine the cycle times of wafer fabs and test facilities, to decide on the modeling of supply uncertainty and validate them against real-world data. Following this, we analyze historical forecast errors. Finally, we conduct experiments to assess how uncertainties in demand and supply affect the OTD and IOH metrics and compare the results with those from deterministic planning.

4.1 Use Case

The front-end supply chain under consideration consists of two in-house wafer fabs (Fab 1 and 2), one outsourcing wafer fab (Fab 3), and two wafer test facilities. Fab 2 and Fab 3 supply to the same wafer test facility, while Fab 1 supplies to the other one. The simulation model includes a total of 8 different wafers. These products were selected to represent one wafer technology per fab, with multiple products modeled within each selected technology. In the results, the performance of wafers C and D from wafer Fabs 1 and 2, respectively, will be reflected. Results are obtained through 500 simulation replications.

4.2 Supply Modeling Validation

First, we determine whether the cycle time is influenced by the workload by fitting a polynomial regression model to the data. We analyze data between the beginning of 2020 to mid 2024. Fab 3 is an outsourced fab, so data on WIP levels could not be collected. The R-squared score of 0.695 for Fab 1 indicates a good fit, whereas the score of 0.106 for Fab 2 suggests a poor fit. Therefore, it can be assumed that wafer Fab 1 has load-dependent cycle times, while wafer Fab 2 does not. S-curves are fitted to the cycle times of Fab 2 and for Fab 3 we determined deterministic cycle times. Figure 2 depicts the validation of arrivals at the die bank for products produced in the different wafer fabs, taking wafer test and yield into account. The predicted cumulative arrivals at the die bank are represented by a line and a shaded region indicating the 95% confidence interval of the observed values. A historical master production plan was used to determine the wafer starts. The simulation model's output is then validated against actual weekly arrival data at the die bank, allowing for a direct comparison between predicted and observed performance. The vertical dotted line represents the total cycle time, calculated by summing all process step durations. Deviations beyond this point may result from changes in the master plan. Since the data of actual arrivals is in weeks, we look at weekly arrivals of the wafers at the die bank. The weeks and quantities are normalized between 0 and 1. The simulation model accurately simulates the supply of wafers as the values fall within the confidence intervals.



(a) Fab 1. (b) Fab 2. (c) Fab 3.

Figure 2: Validation of supply of all wafer fabs.

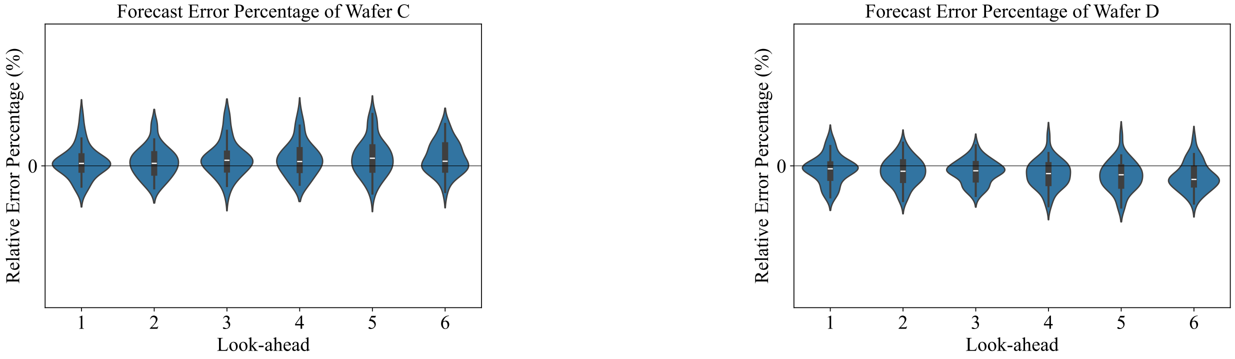
Table 2 denotes the accuracy of the simulation model both with and without the consideration of load-dependent cycle times. Three different master plans from different dates are simulated, and the average MAPE is calculated for the different wafers for each week leading up to the wafer's lead time. The results presented in Table 2 show the average MAPE with a 95% confidence interval. It is evident that including load-dependent cycle times improves simulation model accuracy for the different products, as demonstrated by a better average MAPE in all cases except one, where the results were similar.

Table 2: Difference load-dependent cycle times.

Runweek	Wafer	MAPE with	MAPE without
A	Wafer A	12.18%±1.28%	47.16%±5.23%
B	Wafer A	9.02%±0.86%	40.39%±6.11%
C	Wafer A	27.12%±1.30%	35.60%±0.45%
A	Wafer B	16.64%±1.51%	25.56%±2.10%
B	Wafer B	34.58%±1.74%	34.15%±0.96%
C	Wafer B	12.06%±0.76%	25.97%±1.81%
A	Wafer C	54.08%±2.26%	64.48%±1.26%
B	Wafer C	18.22%±1.20%	29.33%±1.85%
C	Wafer C	17.79%±1.07%	31.87%±1.61%

4.3 Forecast Error Distributions

For all unique sets, we compute the historical forecast percentage errors, see Figure 3. We analyze data between the end of 2020 to mid 2024. The relative forecast error percentage (%) on the y-axis indicates the percentage difference between the forecast and the actual requested demand. A positive forecast error means under-forecasting, while a negative relative forecast error means over-forecasting. The figure indicates that the bias of the forecast percentage error is positive for wafer C and negative for wafer D. This is evident from the fact that the median line inside the boxplot of the violin plot is positioned above zero for positive bias and below zero for negative bias. Moreover, the bias in Figure 3b is larger than in Figure 3a. This is reflected by the larger distance of wafer D's median from zero and the more consistently shifted data points, which indicate a greater deviation from the reference value compared to wafer C. A distribution is fitted to the percentage errors for each combination of subset and look-ahead period. All fitted distributions yielded p-values greater than 0.05, indicating statistically acceptable goodness-of-fit.



(a) Wafer C.

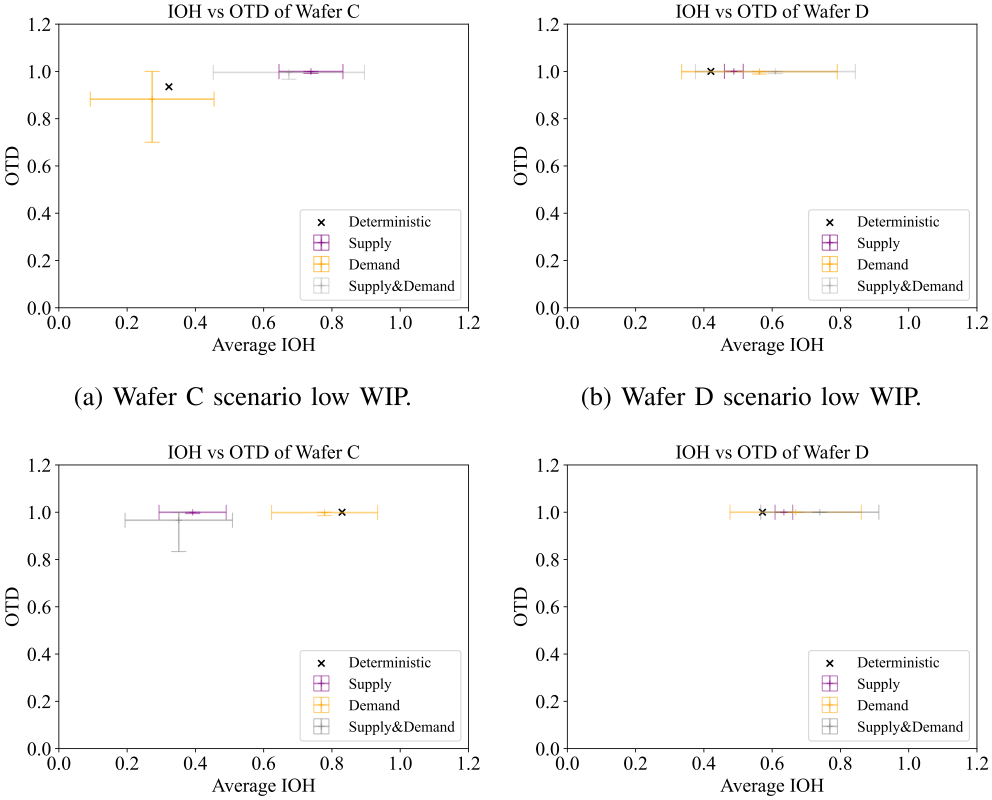
(b) Wafer D.

Figure 3: Historical forecast errors of wafers C and D.

4.4 Results

We utilize our method to assess the impact of supply and demand uncertainty on the performance indicators on-time delivery and inventory on hand at the die bank. This assessment aims to highlight the discrepancy between deterministic planning and the consideration of uncertainty.

To achieve this, two scenarios are defined. Each scenario involves manually adjusting the WIP level of Fab 1. In scenario 1, the low WIP levels of wafer Fab 1 lead to an overestimation of the cycle time and in scenario 2 the High WIP levels an underestimation. We test each scenario under a different combination of uncertainty namely, no uncertainty, only supply uncertainty, only demand uncertainty, and uncertainty in both demand and supply. In each case, the remaining parameters are kept deterministic to isolate the effects of the specified uncertainty. Each scenario uses a different historical master plan, but within a scenario, the same plan is used for all uncertainty simulations. Each combination of uncertainty is simulated over a 26-week horizon. A 26-week horizon offers a balanced time frame that is long enough to capture the impact, yet short enough to support actionable, tactical decision-making. Figure 4 illustrates the OTD and inventory on hand performance for both scenarios and the two wafers. The crosses on the graph represent the mean values with a 95% confidence interval of the observed values. Both OTD and inventory on hand are normalized between 0 and 1. The 'x' signifies the expected performance indicator values when only deterministic parameters are considered. The OTD performance is not affected for wafer D, whereas for wafer C, the uncertainties do impact the expected OTD performance.



(a) Wafer C scenario low WIP.

(b) Wafer D scenario low WIP.

(c) Wafer C scenario high WIP.

(d) Wafer D scenario high WIP.

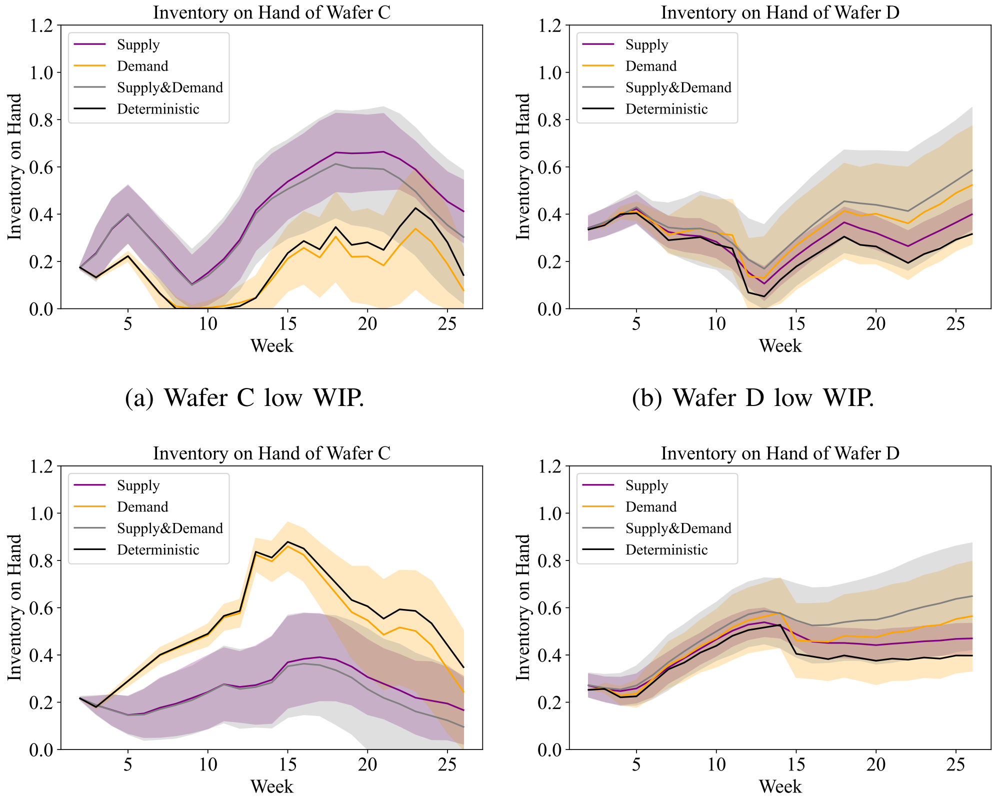
Figure 4: IOH vs OTD for different scenarios.

The warm-up period is two weeks, which allows the simulation model to reach the desired conditions. This period is reasonable because it represents the time required for the first products to arrive at the die bank. WIP levels from the wafer fabrication facilities are available and used to initialize the model, ensuring that it begins in a realistic steady state. As a result, the system is already producing outputs that reflect normal operating behavior. From this point onward, the results are recorded.

When accounting for supply uncertainty, the inventory on hand for wafer D differs minimally from the deterministic results in both scenarios, as the mean inventory on hand is relatively close. Cycle times in both wafer fabrication and wafer test are slightly overestimated, contributing to this small difference.

Additionally, the low standard deviation, shown by the narrow cross, reflects minimal cycle time variation. In contrast, wafer C demonstrates significant differences in inventory on hand in both scenarios with a larger uncertainty, resulting in either more or less inventory being built than anticipated. This even improved the OTD in the low WIP scenario, as can be seen in Figure 4a. The differences when considering only demand uncertainty can be explained by bias. Wafer D's negative bias causes over-forecasting and higher inventory, while Wafer C's positive bias leads to under-forecasting and lower inventory. Considering both supply and demand uncertainties amplifies their effects. Overestimated forecasts and lead times excessive inventory building, while overestimated lead times and underestimated demand partially offset each other. Overall, the uncertainty increases due to compounded uncertainty. For Wafer D, demand uncertainty dominates, whereas Wafer C is affected by both. For wafer C, incorporating uncertainty in both supply and demand improves OTD in the low WIP scenario but worsens OTD in the high WIP scenario. The deterministic results fall outside the confidence intervals for wafer C and are on the edge of the confidence intervals for wafer D.

In Figure 5, we show detailed simulation results where inventory on hands over time are compared for the different scenarios. This provides clear insights into how inventory levels evolve over time, helping to identify potential shortages or overstock situations. For each period, the mean inventory on hand and the standard deviation are calculated with a 95% confidence interval. The inventory on hand is normalized between 0 and 1.



(a) Wafer C low WIP.

(b) Wafer D low WIP.

(c) Wafer C high WIP.

(d) Wafer D high WIP.

Figure 5: IOH for different scenarios.

When only demand uncertainty is considered, the confidence intervals widen, as expected, due to decreasing forecast accuracy over time. This trend continues when both demand and supply uncertainties are factored in, with the intervals expanding and extending. Additionally, in Figures 5b and 5d, the deterministic results are located at the edge of the confidence interval. This indicates that, in the worst-case scenario, the inventory on hand aligns with these deterministic outcomes, suggesting that the current planning approach is conservative. Moreover, in Figure 5c, the deterministic results fall outside the confidence intervals of the line, indicating a significant difference between expected values. Although the average inventory on hand in Figure 5a is higher compared to the deterministic results, the increased uncertainty causes the deterministic values to fall within the confidence intervals for the final weeks. This indicates reduced certainty regarding the distinction between the values.

5 DISCUSSION

Originality. We propose an aggregated simulation model of the front-end semiconductor supply chain. Unlike the detailed simulation models often required in the literature, our approach does not rely on such a model. The simulation model is based on master data structures and applied to the NXP Semiconductors N.V. master data. The simulation modeling of supply is validated using real-world data. To account for supply uncertainty, we utilize discrete distributions of historical cycle times and integrate load-dependent cycle times using a non-linear regression model. Additionally, we incorporate demand uncertainty by employing distributions of historical forecast errors and adding them to future forecasts.

Methodology. Our approach models the front-end supply chain at an aggregated level using master data, avoiding the need for detailed simulation models. This reduces computational effort, simplifies data management, and shortens development time. The proposed aggregation level may have the drawback of losing important details, which could result in less accurate outcomes.

Although we considered clustering to model load-dependent cycle time distributions, it proved less accurate than our chosen method. We assume that both the cycle time distributions and demand forecasting will not improve over time, and therefore, the historical distributions will still be relevant in the future. The limitation of this approach is that the entire fab must be modeled to update the WIP levels, or assumptions must be made regarding the development of these levels. There is a concern about over-fitting past percentage errors onto future demand forecasts. Additionally, the lack of adjustments to future forecasts during the simulation amplifies the effects of growing demand uncertainty over time. If no distribution fits the historical forecast percentage errors, we suggest sampling discrete historical percentage errors.

Application. The proposed simulation model is applied to assess master plans and their performance in the face of demand and supply uncertainty, comparing it to deterministic planning. Real-world data is used to validate the accuracy of the modeling of supply uncertainty. The current approach for modeling supply uncertainty accurately simulates supply for the current set of wafers. However, it could be further validated for various wafer technologies within the different wafer fabs. While other wafer technologies within Fab 1 demonstrated a relationship between WIP levels and expected cycle times, this relationship has not yet been validated.

6 MANAGERIAL INSIGHTS

The proposed aggregated simulation approach enables practitioners to validate master plans under uncertain supply and demand conditions. Additionally, practitioners can adjust the master plan and validate it under uncertainty rather than solely deterministic parameters. The simulation model provides insights into how the uncertainties impact performance indicators, helping practitioners assess both the expected average performance and the likelihood of meeting performance targets. The results demonstrate that biases in the parameters could significantly impact the performance indicators, highlighting the added value of the simulation. These capabilities enhance the ability to make informed decisions in master planning, balancing inventory costs with On-Time-Delivery.

Although our study primarily focuses on master plan validation, our proposed simulation model has other compelling applications, such as incident analysis or high level capacity planning. This is because the aggregation level of the simulation reduces the computation time, allowing for the testing of multiple scenarios. By only modeling key bottleneck resources aligned with the company's existing master data structure, we not only reduce the time needed to develop the simulation model but also lessen the computational burden compared to more detailed approaches. This enables the model to efficiently test various scenarios, thereby assisting practitioners in making strategic decisions.

7 CONCLUSION

In this study, we present an aggregated simulation model of the front-end semiconductor supply chain. We demonstrate our proposed method by analyzing master plans and deviations in the expected performance due to uncertainty. To account for future supply and demand uncertainties, we use historical data on cycle times, forecast errors, and future forecasts. Upon analyzing the cycle time distributions across different wafer fabs, we observed a correlation between cycle times and work-in-progress (WIP), indicating load-dependent cycle times for one wafer fab. The approach of modeling supply is validated using real-world data and yields accurate results. Our simulation results highlight the impact of demand or supply uncertainty on inventory levels and On-Time Delivery performance. Specifically, we found that the presence of a large enough bias renders deterministic outcomes insignificant. Moreover, when both supply and demand are taken into account together, the standard deviation (uncertainty) increases compared to when they are considered separately.

In future research, it would be of great interest to expand the study by incorporating a wider range of products. Additionally, it would be beneficial to integrate a simulation model of the back-end supply chain to factor in the capacity constraints of the back-end operations. Moreover, it would be interesting to extend the simulation by integrating forecast evolution, enabling the model to be used in a rolling horizon. Instead of just analyzing the work-in-progress (WIP) of the entire fab as done in the current method, future research could also consider the WIP of the wafer technology. Given that this study's findings are based on data from a single company, it is crucial for future research to consider using the same approach with data from various companies to confirm the results throughout the semiconductor industry.

REFERENCES

Atherton, L. F., and R. W. Atherton. 1995. *Wafer fabrication: Factory performance and analysis*, Volume 339. Boston: Springer Science & Business Media.

Chien, C.-F., S. Dauzere-Peres, H. Ehm, J. W. Fowler, Z. Jiang, S. Krishnaswamy, *et al*. 2008. "Modeling and Analysis of Semiconductor Manufacturing in a Shrinking World: Challenges and Successes". In *2008 Winter Simulation Conference (WSC)*, 2093–2099 https://doi.org/10.1109/WSC.2008.4736306.

Curry, G., and R. Feldman. 2011. *Manufacturing Systems Modeling and Analysis*. Berlin: Springer.

Deenen, P. C., J. Middelhuis, A. Akcay, and I. Adan. 2024. "Data-Driven Aggregate Modeling of a Semiconductor Wafer Fab to Predict WIP Levels and Cycle Time Distributions". *Flexible Services and Manufacturing Journal* 36:567–596.

Duarte, B., J. Fowler, K. Knutson, E. Gel, and D. Shunk. 2007. "A Compact Abstraction of Manufacturing Nodes in a Supply Network". *International Journal of Simulation and Process Modelling* 3.

Ewen, H., L. Mönch, H. Ehm, T. Ponsignon, J. W. Fowler, and L. Forstner. 2017. "A Testbed for Simulating Semiconductor Supply Chains". *IEEE Transactions on Semiconductor Manufacturing* 30(3):293–305.

Fowler, J., L. Mönch, and T. Ponsignon. 2015. "Discrete-Event Simulation for Semiconductor Wafer Fabrication Facilities: A Tutorial". *The International Journal of Industrial Engineering: Theory, Applications and Practice* 22:661–682.

Hartwick, A., A. Ismail, B. K. Valladão Novais, M. Zeeshan, and H. Ehm. 2023. "System Dynamics Simulation of External Supply Chain Disruptions on a Simplified Semiconductor Supply Chain". In *2023 Winter Simulation Conference (WSC)*, 863–874 https://doi.org/10.1109/WSC60868.2023.10408593.

Hung, Y.-F., and C.-B. Chang. 1999. "Determining Safety Stocks for Production Planning in Uncertain Manufacturing". *International Journal of Production Economics* 58(2):199–208.

Kopp, D., M. Hassoun, A. Kalir, and L. Mönch. 2020. "SMT2020—A Semiconductor Manufacturing Testbed". *IEEE Transactions on Semiconductor Manufacturing* 33(4):522–531.

Leachman, R. 2002. "Semiconductor Production Planning". In *Handbook of Applied Optimization*, 746–763. New York: Oxford University Press.

Morrice, D., R. Valdez, J. Chida, and M. Eido. 2005. "Discrete Event Simulation in Supply Chain Planning and Inventory Control at Freescale Semiconductor Inc". In *2005 Winter Simulation Conference (WSC)*, 1718–1724 https://doi.org/10.1109/WSC.2005.1574444.

Mönch, L., J. Fowler, and S. Mason. 2013. *Production Planning and Control for Semiconductor Wafer Fabrication Facilities: Modeling, Analysis, and Systems*. New York: Springer.

Mönch, L., R. Uzsoy, and J. Fowler. 2018. "A Survey of Semiconductor Supply Chain Models Part III: Master Planning, Production Planning, and Demand Fulfilment". *International Journal of Production Research* 56(13):4565–4584.

Orcun, S., and R. Uzsoy. 2011. *The Effects of Production Planning on the Dynamic Behavior of a Simple Supply Chain: An Experimental Study*, 43–80. New York, NY: Springer New York.

Ponsignon, T., and L. Mönch. 2014. "Simulation-Based Performance Assessment of Master Planning Approaches in Semiconductor Manufacturing". *Omega* 46:21–35.

Rose, O. 1999. "Estimation of the Cycle Time Distribution of a Wafer Fab by a Simple Simulation Model". In *Proceedings of the International Conference on Semiconductor Manufacturing Operational Modeling and Simulation*, 133–138.

Rose, O. 2007a. *Benefits and Drawbacks of Simple Models for Complex Production Systems*, Volume 2008, 91–118. Springer Berlin Heidelberg.

Rose, O. 2007b. "Improved Simple Simulation Models for Semiconductor Wafer Factories". In *2007 Winter Simulation Conference (WSC)*, 1708–1712 https://doi.org/10.1109/WSC.2007.4419793.

Rosman, C., E. Weijers, K. Schelthoff, W. van Jaarsveld, A. Akcay, and I. Adan. 2024. "Aggregated Simulation Modeling to Assess Product-Specific Safety Stock Targets During Market Up- and Downswings: A Case Study". In *2024 Winter Simulation Conference (WSC)*, 1931–1942 https://doi.org/10.1109/WSC63780.2024.10838984.

Shanthikumar, J., S. Ding, and M. Zhang. 2007. "Queueing Theory for Semiconductor Manufacturing Systems: A Survey and Open Problems". *EEE Transactions on Automation Science and Engineering* 4:513–522.

Spearman, M. 2014. "Of Physics and Factory Physics". *Production and Operations Management* 23:1875–1885.

Wang, C.-H., and J.-Y. Chen. 2019. "Demand Forecasting and Financial Estimation Considering the Interactive Dynamics of Semiconductor Supply-chain Companies". *Computers Industrial Engineering* 138:106104.

Wu, K. 2005. "An Examination of Variability and its Basic Properties for a Factory". *IEEE Transactions on Semiconductor Manufacturing* 18(1):214–221.

Ziarnetzky, T., L. Mönch, and R. Uzsoy. 2020. "Simulation-Based Performance Assessment of Production Planning Models With Safety Stock and Forecast Evolution in Semiconductor Wafer Fabrication". *IEEE Transactions on Semiconductor Manufacturing* 33(1):1–12.

AUTHOR BIOGRAPHIES

AARON SIEDERS is an analyst in the Department of Sales Operations at NXP Semiconductors N.V. He obtained his master's degree in Operations Management & Logistics at the Eindhoven University of Technology. His email address is aaron.sieders@nxp.com.

CAS ROSMAN is a doctoral candidate in the Department of Planning IT at NXP Semiconductors N.V. and in the Department of Industrial Engineering and Innovation Sciences at the Eindhoven University of Technology. He obtained his master's degree in Operations Management Logistics at the Eindhoven University of Technology. His primary research interest is in the area of simulation modeling and decentralized decision-making in supply chain management. His email address is cas.rosman@nxp.com.

COLLIN DRENT is an Assistant Professor of Operations Management in the School of Industrial Engineering at Eindhoven University of Technology. He received his Ph.D. in Applied Mathematics from the same university. His research focuses on optimal decision-making under uncertainty — leveraging operations research, statistics, and machine learning. His work has applications in manufacturing, maintenance and healthcare operations. His email address is c.drent@tue.nl

ALP AKCAY is an Associate Professor in the Department of Mechanical and Industrial Engineering at Northeastern University. He received his Ph.D. in Operations Management and Manufacturing from Carnegie Mellon University. His research focuses on the planning and control of manufacturing systems and supply chains, using techniques from stochastic operations research, machine learning, and simulation. His email address is a.akcay@northeastern.edu.