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.

IMPROVING PLAN STABILITY IN SEMICONDUCTOR MANUFACTURING THROUGH STOCHASTIC OPTIMIZATION: A CASE STUDY

Eric Weijers[1], Nino Sluijter[1], Gijs Hogers[1], Kai Schelthoff[1], Ivo Adan[2], and Willem van Jaarsveld[2]

[1]Dept. of Supply Chain Innovation, NXP Semiconductors N.V., Eindhoven, NETHERLANDS
[2]Dept. of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology,
Eindhoven, NETHERLANDS

ABSTRACT

In this study, we propose a two-stage stochastic programming method to improve plan stability in semiconductor supply chain master planning in a rolling horizon setting. The two-stage programming model is applied to real-world data from NXP Semiconductors N.V. to assess the quality of generated plans based on the KPIs plan stability, on-time delivery, and inventory position. We also compare the performance of two-stage stochastic programming to linear programming. To model demand uncertainty, we propose to fit distributions to historical demand data from which stochastic demand can be sampled. For modeling supply, we propose an aggregated rolling horizon simulation model of the front-end supply chain. Based on the performed experiments, we conclude that two-stage programming outperforms LP in terms of plan stability, while performing comparably in terms of inventory position and on-time delivery.

1 INTRODUCTION

The semiconductor supply chain consists of a front-end and a back-end process. In the front-end process, electronic circuits are patterned onto raw wafers. In the back-end process, the patterned wafers are cut into individual dies, which are then assembled, tested, and packaged.

Production of patterned wafers in the front-end process typically has a lead time between 10-15 weeks (Mönch et al. 2018), whereas processing wafers to dies in the back-end process typically has a lead time between 1-2 weeks. Because of this difference, both processes use different manufacturing paradigms. In front-end manufacturing, a Make To Stock (MTS) paradigm is used, and production orders start mainly based on forecasted demand. In the back-end manufacturing process, a Make To Order (MTO) paradigm is used, and production orders start based on incoming customer orders. A die bank decouples both processes. This is an inventory point where patterned wafers coming from the front-end are stored until they are needed in the back-end.

In semiconductor supply chains, planning the wafer starts in the front end is crucial to maintain sufficient inventory levels at the die bank to cover customer orders pulling wafers from the die bank while avoiding excess inventory, which may lead to obsolescence. The goal of front-end master planning is to optimally coordinate demand and supply based on input parameters representing the most recent information. The optimality of a plan is based on On Time Delivery (OTD) and inventory costs. For optimality, OTD should be maximized, while inventory costs should be minimized. As these are conflicting objectives, master planning strives to find a balance between the two. The input parameters to the plan consist of demand, inventory, resource, and wafer parameters. Due to both demand and supply uncertainty, plans must be periodically optimized. Currently, Linear Programming (LP) models with deterministic parameters are used to optimize the front-end plan. Optimization is done weekly with a two-year planning horizon for strategic decision-making purposes. A plan contains the weekly allocation of wafers to different factories. Shop-floor planning is done by the factories themselves. When analyzing week-to-week plans, it can be noted that the plan must frequently change to remain optimal under the newly observed (deterministic) input

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

parameters. As the stochastic nature of the input parameters is not considered during plan optimization, the plan is not robust to changes in input parameters over time. The main drivers of demand uncertainty are customer behavior (delaying or expediting orders), market cyclicality, and forecast accuracy. The latter is important in wafer planning because of the MTS manufacturing paradigm. Supply uncertainty arises from capacity changes, yield changes, and quality incidents. Similar to the demand uncertainty, it is currently not considered during plan optimization. In addition to the demand and supply uncertainty, the nature of LP affects plan robustness. LP optimizes towards extreme point solutions. Additionally, in the context of highly degenerated planning problems, often many alternative optimal solutions exist. Both properties contribute to the tendency of LP to produce variable plans over time, even when input parameters are fairly stable (Mönch et al. 2018).

To make the supply chain master planning more robust to stochasticity in the input parameters, a method is needed that can find near-optimal solutions while accounting for the stochastic nature of the input parameters. In this paper, we present a method based on two-stage stochastic programming (2SP) in a rolling horizon setting to reduce plan nervousness, considering demand uncertainty. The proposed method is demonstrated using a case study with real-life data from NXP Semiconductors N.V. To the best of our knowledge, this is the first application of 2SP to reduce plan nervousness in semiconductor supply chain master planning. After discussing relevant literature in section 2, we propose our method in section 3. The problem is mathematically defined in section 4. We apply the method to a real-life case study in section 5. the paper is concluded with a discussion in section 6, and conclusions in section 7.

2 LITERATURE REVIEW

De Kok and Inderfurth (1997) discuss the need to define plan stability as an independent KPI next to the traditional performance measures when analyzing a rolling horizon planning framework. Several plan stability measures exist in literature (Kabak and Ornek 2009). In the literature, there are mainly two approaches to address the problem of plan nervousness under stochasticity. Reactive (or static) approaches strive to find a stationary state of the system. Proactive (or dynamic) approaches strive to identify and dampen the root cause of plan nervousness. In this literature review, we will focus on papers discussing reactive (or static) approaches in a rolling horizon setting.

Kimms (1998) propose an iterative procedure in which plan instability is a variable in the mathematical formulation of the planning problem, and a constraint is added to enforce an upper bound on the instability of the plan. This study shows the benefits of accounting for plan stability during plan calculation. Ponsignon and Mönch (2014) combine plan freezing with two reactive approaches for plan computation, Genetic Algorithm (GA) and a rule-based assignment (RA) procedure. They conclude that RA is the better approach for generating stable plans and GA for generating plans with low production costs. Ziarnetzky et al. (2019) extend their study on chance constraint programming with a study of the effect of freezing on the nervousness of production planning in wafer fabrication using a rolling horizon setting. They conclude that incorporating frozen periods in the planning model improves plan stability but at the cost of expected profit. Similar to this study, Sáez et al. (2023) propose a combination of a mathematical model and intelligent agents. The plan generated by the mathematical model serves as a baseline for the decentralized agents to manage the nervousness of the plan. The authors conclude that the agents are capable of producing more stable plans while only slightly increasing production costs.

A large body of papers focuses on chance constraint programming. We highlight the two papers most relevant to this study. Lin and Uzsoy (2016) define two chance constraint production planning models specifically developed for systems with stochastic demand in a rolling horizon setting. The application of chance constraints results in building safety stock to mitigate changes in demand and thus results in a more stable plan. Ziarnetzky et al. (2018) focus on improving planning performance through exploiting updated demand information through forecast evolution. Similar to the latter research, the authors address plan nervousness by incorporating chance constraints in combination with safety stocks. In addition, the proposed stability measure makes sure that changes earlier in the plan are penalized more heavily than changes later in the plan. They conclude that considering forecast evolution leads to improved performance as long as some excess capacity is available.

Another widely used approach is to incorporate plan stability in the objective function of the planning problem. Koca et al. (2018) define a nervousness cost function and solve a second-order cone mixed integer program (SOCMIP) to create stable plans. Results indicate that the nervousness costs are significantly reduced, and more stable production plans are generated with only a relatively small increase in total cost. Özelkan et al. (2023) propose a bi-objective aggregate production planning model where plan stability is incorporated as a second objective alongside the traditional cost objective. The proposed method is tested on multiple industry-based use cases and is proven to outperform traditional aggregate production planning models concerning plan nervousness while yielding comparable production costs. Friese and Helber (2023) define an interactive multi-objective optimization algorithm. In each decision stage, a Pareto optimal solution set is generated, which supports the decision-maker in making the tradeoff between plan nervousness and production costs. Although the performance of the method seems good, handling larger (real-life) problem instances remains future research.

Lastly, we highlight the paper of Rashidi et al. (2024) in which the semiconductor master planning under demand and yield uncertainty is solved using a three-stage stochastic programming model. While the focus of the paper is on maximizing expected profit, the authors emphasize the potential of stochastic programming to include other objectives, such as plan stability.

We conclude that plan stability is recognized as an important KPI when considering (master) planning in a rolling horizon setting. Although an increase in research on dampening strategies is visible in recent years, few studies focus on incorporating uncertainty into the planning model. Chance-constrained programming has emerged as a promising direction, particularly for managing stochastic demand. However, this method often requires predefined confidence levels and rigid constraint formulations, which can limit its adaptability in environments with evolving or poorly characterized uncertainty. In contrast, two-stage stochastic programming offers a more flexible framework by modeling uncertainty through scenario-based representations and allowing for a broader range of responses to different outcomes. Therefore, this study, which investigates a simple 2SP method in a rolling horizon setting, is original, practical, and applicable to supply chain planning problems.

3 METHOD

We propose a 2SP model to reduce plan nervousness in semiconductor supply chain master planning in a rolling horizon setting. This section consists of four parts. First, we propose to model demand uncertainty using demand distributions fitted to historical data. Second, we propose a rolling horizon aggregated simulation model to compare plans based on plan stability, OTD, and inventory position. Third, to compare plans based on plan stability, we propose a stability measure. Finally, we propose a 2SP to reduce plan nervousness compared to LP.

3.1 Demand Modeling

To generate stochastic demand, a distribution is fitted to the historical demand data of a product, and sampling from this distribution is used to generate the demand. We focus on high-demand products (products in ramp-up or ramp-down are not considered) and distinguish between four demand patterns using the demand pattern classification framework presented by Syntetos et al. (2005). The four patterns are intermittent, lumpy, smooth, and erratic.

Intermittent and lumpy demand patterns are characterized by alternating periods of demand and one or more periods without demand, while lumpy patterns are characterized by a higher variability in demand sizes. Both the arrival of demand and its magnitude must be modeled. For modeling demand arrivals, we utilize the Markov Chain (MC) approach proposed by Prak and Rogetzer (2022). In this approach, each state in the MC represents the number of periods since the last demand occurred. With probability p we transition back to the initial state and a new demand occurs, and with probability $1-p$ we transition to the next state without a demand occurrence. The transition probability p for each state is estimated from the demand data. To model demand sizes, we utilize the widely used Gamma distribution (Turrini and Meissner 2019). Although it is a continuous distribution, it can only approximate the discrete demand distribution. The large demand volumes of the high-demand products within the use case mitigate this effect.

Smooth and erratic demand patterns are closely related, as erratic patterns are similar to smooth patterns but with greater variability in demand sizes. Initial data analysis of smooth and erratic demand shows that periods with zero demand are present in the data, though less frequently and for shorter periods compared with intermittent and lumpy demand. We propose to use the same approach as used for intermittent and lumpy demand.

3.1.1 Rolling Horizon Implementation

The demand generation methods discussed above are used to generate a base demand signal for the rolling horizon. Per product, for the complete planning horizon, a demand from the generated distributions is sampled to create the base demand signal. In this research, we use a planning horizon of 26 weeks. The base demand is then updated over time to create a rolling horizon environment where demand information is updated at the beginning of each planning period. Based on the historical demand data per product, a probability that the demand will deviate is derived. To determine the amplitude of the demand change, a normally distributed random variate $R \sim N(0, \sigma^2)$ is used as presented by Ponsignon and Mönch (2014). We define the standard deviation of R as $3\sigma = 0.1$. As such, most of the demand deviations will be within 10%. As the demand is updated at the beginning of each period, the plan must be reoptimized each period. The decisions for the current period, as indicated by the reoptimized plan, are implemented.

3.2 Supply Modeling

The simulation model enables measuring KPI performance (OTD and inventory position) and comparison of different plan generation methods (LP and 2SP). The aggregated Discrete Event simulation (DES) method presented by Rosman et al. (2024) is used to build a simulation model of the front-end supply chain. The main advantages of this method are (1) the model is based on existing master data structures, reducing the required data maintenance effort, and (2) the aggregation level of the model limits computational expense compared to more detailed models. The front-end supply chain consists of two main steps. Firstly, wafer fabrication in which electronic circuits are patterned on the wafer. Secondly, wafer test, in which the quality of the wafer is checked. Between both steps, the wafer is moved from the fabrication facility to the test facility, incurring transportation time.

3.3 Stability Measure

To define plan stability between two consecutive plans, we modify the stability measure presented by Kabak and Ornek (2009). The measure is reformulated to reflect the product-oriented structure of wafer planning:

$$I^Q(k) = \frac{\sum_{p=1}^{P} \sum_{t=0}^{T-1} \left| Q_{pt}^k - Q_{pt}^{k-1} \right| \cdot W^Q(t)}{\Delta q}.$$

Q_{pt}^k and Q_{pt}^{k-1} are the planned order quantities for product p at time t in planning cycles k and $k-1$, respectively. $W^Q(t)$ is a decreasing weight function that emphasizes the importance of changes occurring in earlier time periods of the planning horizon and is defined as $W^Q(t) = At^{-B}$. A and B are positive constants that control the scale and steepness of the decay. Both values are set according to the findings presented by Kabak and Ornek (2009), where $A = 1.5$ and $B = 1.2$. Δq is a global normalization factor capturing the maximum weighted deviation in planned order quantities across all products and planning cycles:

$$\Delta q = \max_k \sum_{p=1}^{P} \sum_{t=0}^{T-1} \left| Q_{pt}^k - Q_{pt}^{k-1} \right| \cdot W^Q(t).$$

To compare the stability of multiple models, in our case, LP and 2SP, the normalization factor should be generalized over the different models. This is achieved through defining the global normalization factor as the maximum of the individual factors across the different models. Combining the above, the overall stability measure across all planning cycles k is then formulated as follows:

$$S^Q = 1 - \frac{1}{K} \sum_{k=1}^{K} I^Q(k).$$

Concluding, the proposed measure accounts for the number of changes between two consecutive plans, the quantity of the changes, and the timing, where changes in earlier periods result in a lower stability score than changes in later periods. A higher score means a higher stability.

3.4 Two-stage Stochastic Programming Model

The main strength of 2SP is that more information on the demand distribution is provided to the model compared with the deterministic input parameters used by LP. This is done by introducing multiple demand scenarios, where each scenario occurs with a certain probability. 2SP then strives to optimize the expected value of the objective function across all scenarios. If a demand change occurs that falls within one of the scenarios used for optimization, the plan doesn't have to be changed. To define the base demand for a scenario, we use the demand distributions presented in subsection 3.1. For each scenario, demands for each product and each time period are sampled. One scenario will be equal to the scenario provided to LP. In the current research, we consider 5 demand scenarios for 2SP. As the demand in all scenarios is randomly sampled from the demand distributions, we assign equal probabilities of 0.2. As this is a first exploratory study into using 2SP to improve plan stability, the effect of the number of scenarios and the associated probabilities has not yet been researched. This is highly recommended for future research, as literature shows that selecting the number of scenarios and the associated probabilities is a non-trivial task.

4 MATHEMATICAL FORMULATION

Based on several input parameters, the wafer plan strives to maximize OTD while minimizing inventory costs, where OTD is the highest priority objective. During optimization, for each time period t (week) in set T, for all products p in set P, we strive to plan all confirmed and forecasted demand D_{pt}^{ϑ}, where $\vartheta \in \{d, f\}$. Confirmed demand will be prioritized over forecasted demand using layered objective levels. Similarly, forecasted demand will be prioritized over safety stock targets and pre-building. We assume that each product p (wafer) has one supply chain to be produced. A supply chain is defined as the link between two stock points. In wafer fabrication, these are the raw material inventory and the die bank. When demand is allocated to a supply chain, allocation and lateness costs are incurred. For confirmed and forecasted demand, these costs are denoted with C_{pt}^{ϑ}, where $\vartheta \in \{d, f\}$. The allocation and lateness costs include a time component to penalize allocation after the deadline t_d. When not all demand can be allocated due to limited capacity, a penalty cost is incurred and is denoted with N_p^{ϑ}, where $\vartheta \in \{d, f\}$. If, after allocating all confirmed and forecasted demand, capacity is still available and a target safety stock S_{pt} is set for product p, safety stock is planned. If, after allocating all demand and safety stock, capacity is still available and the target capacity Tar_{rt} for resource r at time t is not met, products from set $\hat{P}$, which are eligible for pre-building, are planned. Capacities are given per resource r in set R per time period t, and are denoted with Cap_{rt}. All sets, parameters, and variables used to define the wafer planning problem are presented in Table 1. We will continue with mathematically defining the wafer planning problem as an LP model.

Table 1: Sets, parameters, and variables used to define the wafer planning problem

Indices and Sets:	
$p \in P$	Products index and set, representing different types of products available for allocation
$\hat{P} \subseteq P$	A subset of products that can be used for pre-build
$r \in R$	Resources index and set, denoting the different resources or machines used in production
$t \in T$	Time periods index and set (in weeks)
ϑ	A set containing the indices d and f, indicating confirmed or forecasted demand
Parameters:	
D^d_{pt}	Confirmed demand d for product p at time t
D^f_{pt}	Forecasted demand f for product p at time t
S_{pt}	Target safety stock for product p at time t
L_p	The lead time of product p
$t_d \in t$	The deadline by which demand must be fulfilled
U_{pr}	The usage rate of product p when allocated to resource r
Cap_{rt}	Available capacity on resource r at time t
Tar_{rt}	Target capacity on resource r at time t
H^{ϑ}_p	Criticality cost of demand $\vartheta \in \{d, f\}$ for product p, indicating the importance of the demand type and the product
C^{ϑ}_{pt}	Unit allocation costs of demand $\vartheta \in \{d, f\}$ for product p at time t
C^s_{pt}	Unit allocation costs of safety stock for product p at time t
C^b_{pt}	Unit allocation costs of pre-build for product p at time t
N^{ϑ}_p	The costs of not allocating one unit of demand $\vartheta \in \{d, f\}$ for product p
Decision Variables:	
$x^{\vartheta}_{pt} \geq 0$	The quantity of allocated demand $\vartheta \in \{d, f\}$ for product p produced at time t
$v_{pt} \geq 0$	The quantity of allocated safety stock for product p produced at time t
$w_{pt} \geq 0$	The quantity of allocated pre-builds for product p produced at time t
$A^{\vartheta}_p \geq 0$	The quantity of demand $\vartheta \in \{d, f\}$ for product p that is not allocated over the planning horizon
$I_{pt} \geq 0$	Starting inventory position at the die bank for product p at time t
$y_{pt} \geq 0$	The quantity of product p that arrives at I_{pt} at time t
$z^{\vartheta}_{pt} \in \{0, 1\}$	Auxiliary variable for demand $\vartheta \in \{d, f\}$ for product p at time t

4.1 LP Objective

Most real-life problems involve multiple (conflicting) objectives. This also applies to the wafer planning at NXP. To cope with multiple objectives, different objective levels are defined. The different objectives in the objective level set are ordered based on importance. The LP is solved consecutively for each of the objectives, referred to as a layered approach. This process continues until either the last objective is optimized or until the solution space is narrowed down to only one solution. In this formulation, the LP is solved for (1) confirmed demand, (2) forecasted demand, (3) safety stock, and finally (4) pre-builds. These layers are reflected in the objective functions:

$$\text{Objective layer 1:} \quad \min \sum_{p \in P} \sum_{\vartheta \in \{d\}} \left(\sum_{t \in T} (C^{\vartheta}_{pt} \cdot x^{\vartheta}_{pt}) + N^{\vartheta}_p \cdot A^{\vartheta}_p \right),$$

$$\text{Objective layer 2:}\quad \min \sum_{p\in P}\sum_{\vartheta\in\{f\}}\left(\sum_{t\in T}(C_{pt}^{\vartheta}\cdot x_{pt}^{\vartheta})+N_p^{\vartheta}\cdot A_p^{\vartheta}\right),$$

$$\text{Objective layer 3:}\quad \min \sum_{p\in P}\sum_{t\in T} C_{pt}^{s}\cdot v_{pt},$$

$$\text{Objective layer 4:}\quad \min \sum_{p\in P}\sum_{t\in T} C_{pt}^{b}\cdot w_{pt}.$$

4.2 LP Constraints

Criticality Cost. When demand is produced after its due date, lateness costs proportional to its criticality are incurred. We first define a supporting constraint using auxiliary variable $z_{pt}^{\vartheta}\in\{0,1\}$ and an arbitrary large number M:

$$D_{pt}^{\vartheta}\leq Mz_{pt}^{\vartheta}\quad \forall\ \vartheta\in\{d,f\},\ p\in P,\ t\in T.$$

The lateness cost can then be defined:

$$C_{pt}^{\vartheta}=z_{pt}^{\vartheta}H_p^{\vartheta}(t-t_d)\quad \forall\ \vartheta\in\{d,f\},\ p\in P,\ t\in T \text{ if } t>t_d.$$

Demand Fulfillment. Defines that the quantity of demand for product p that is not allocated over the planning horizon equals the total demand for that product over the planning horizon (confirmed demand plus forecast demand) minus the allocated quantity for that product over the planning horizon:

$$A_p^{\vartheta}=\sum_{t\in T}\left(D_{pt}^{\vartheta}-x_{pt}^{\vartheta}\right),\quad \forall\ \vartheta\in\{d,f\},\ p\in P.$$

Ensures that the realized safety stock level for each product p at time t is equal to or below the safety stock target level S_{pt}:

$$v_{pt}\leq S_{pt},\quad \forall\ p\in P,\ t\in T.$$

Ensures that only products in set $\hat{P}$ can be used for pre-building:

$$w_{pt}=0,\quad \forall\ p\notin\hat{P},\ t\in T.$$

Resource Capacity. Ensures that the allocation on any resource at any time does not exceed the available capacity Cap_{rt}:

$$\sum_{p\in P}\left(\sum_{\vartheta\in\{d,f\}}(x_{pt}^{\vartheta})+v_{pt}+w_{pt}\right)U_{pr}\leq \text{Cap}_{rt},\quad \forall\ r\in R,\ t\in T.$$

Ensures that the allocation on any resource at any time exceeds the target capacity Tar_{rt}:

$$\sum_{p\in P}\left(\sum_{\vartheta\in\{d,f\}}(x_{pt}^{\vartheta})+v_{pt}+w_{pt}\right)U_{pr}\geq \text{Tar}_{rt},\quad \forall\ r\in R,\ t\in T.$$

Rolling Horizon. At the end of period t, the starting inventory for the next period $I_{p,t+1}$ is calculated. I_{pt} is defined as the staring inventory for product p at time t at the die bank, $D_{pt}^{\vartheta}\in\{d,f\}$ as the demand for product p at time t that is consumed from the die bank, and y_{pt} as the allocated demand that will arrive at the die bank in period t:

$$I_{p,t+1} = I_{pt} + y_{pt} - \sum_{\vartheta \in \{d,f\}} D^{\vartheta}_{pt}, \quad \forall\ p \in P,\ t \in T.$$

The quantity of demand for product p that arrives at the die bank at time t is defined as the sum of all planned products (confirmed demand, forecasted demand, safety stock, pre-build) that have a lead time L_p such that the products arrive at the die bank at the start of time t:

$$y_{pt} = \sum_{\vartheta \in \{d,f\}} (x^{\vartheta}_{p,t-L_p}) + v_{p,t-L_p} + w_{p,t-L_p}, \quad \forall\ p \in P, t \in T \text{ if } t \geq L_p.$$

Other. Non-negativity constraints for all decision variables:

$$x^{\vartheta}_{pt},\ v_{pt},\ w_{pt},\ I_{pt},\ y_{pt} \geq 0, \quad \forall\ \vartheta \in \{d,f\},\ p \in P,\ t \in T.$$

4.3 Two-stage Stochastic Programming Model

The LP model presented above can be converted to a 2SP model by introducing a set of scenarios, where each scenario occurs with a certain probability p_s. An overview of added or changed sets, parameters, and variables is presented in Table 2. Furthermore, p_s is added to the objective functions.

Table 2: Sets, parameters, and variables that are added or changed to convert LP to 2SP

Indices and Sets:	
$s \in S$	Scenarios index and set, representing different scenarios for parameter realizations
Parameters:	
p_s	The probability that scenario s will occur
D^d_{pts}	Confirmed demand d for product p at time t in scenario s
D^f_{pts}	Forecasted demand f for product p at time t in scenario s
C^{ϑ}_{pts}	Unit allocation costs of demand $\vartheta \in \{d,f\}$ for product p at time t in scenario s
Decision Variables:	
$x^{\vartheta}_{pts} \geq 0$	The quantity of allocated demand $\vartheta \in \{d,f\}$ for product p at time t in scenario s
$v_{pts} \geq 0$	The quantity of allocated safety stock for product p at time t in scenario s
$w_{pts} \geq 0$	The quantity of allocated pre-builds for product p at time t in scenario s
$A^{\vartheta}_{ps} \geq 0$	The quantity of demand $\vartheta \in \{d,f\}$ for product p that is not allocated over the planning horizon in scenario s
$z^{\vartheta}_{pts} \in \{0,1\}$	Auxiliary variable for demand $\vartheta \in \{d,f\}$ for product p at time t in scenario s
$I_{pts} \geq 0$	Starting inventory position at the die bank for product p at time t in scenario s
$y_{pts} \geq 0$	The quantity of product p that arrives at I_{pts} at time t in scenario s

$$\text{Objective layer 1:} \quad \min \sum_{s \in S} \sum_{p \in P} \sum_{\vartheta \in \{d\}} p_s \left(\sum_{t \in T} (C^{\vartheta}_{pts} \cdot x^{\vartheta}_{pts}) + N^{\vartheta}_{p} \cdot A^{\vartheta}_{ps} \right),$$

$$\text{Objective layer 2:} \quad \min \sum_{s \in S} \sum_{p \in P} \sum_{\vartheta \in \{f\}} p_s \left(\sum_{t \in T} (C^{\vartheta}_{pts} \cdot x^{\vartheta}_{pts}) + N^{\vartheta}_{p} \cdot A^{\vartheta}_{ps} \right),$$

$$\text{Objective layer 3:} \quad \min \sum_{s \in S} \sum_{p \in P} \sum_{t \in T} C^{s}_{pt} \cdot v_{pts},$$

$$\text{Objective layer 4:} \quad \min \sum_{s \in S} \sum_{p \in P} \sum_{t \in T} C^{b}_{pt} \cdot w_{pts}.$$

5 APPLICATION

In this section, we apply the proposed methodology from section 3 to a group of products from NXP Semiconductors N.V. All products originate from the same product family, and the dataset only contains high-demand products but with different demand patterns (smooth, erratic, intermittent, and lumpy). We study a real-world scenario in which LP and a two-stage stochastic programming model are used to generate wafer plans. The plans generated by both methods are compared using the KPIs plan stability, OTD, and inventory position.

5.1 Demand Modeling Results

In subsection 3.1, we present a method to fit demand distributions to the historic demand data of a product. For intermittent and lumpy demand patterns, an MC approach is used to model demand arrivals, and a Gamma distribution is used to model the demand sizes. Using this approach, we can fit distributions for 85% of the high-demand products in the use case.

As initial data analysis of smooth and erratic demand showed that periods with zero demand are present in the data, we propose to use the same approach as used for intermittent and lumpy demand. Using this approach, we can fit distributions for 83% of the products in the use case. To test a distribution's goodness-of-fit, we use the Kolmogorov-Smirnov (K-S) test as it is distribution-free and can thus be used for the various distributions discussed above (Syntetos et al. 2012).

5.2 Simulation Results

In this section, we compare the performance of LP and (2SP) when generating wafer plans under demand uncertainty. The quality of both methods is assessed using the KPIs plan stability, OTD, and inventory position. To define plan stability, we use the stability measure presented in subsection 3.3. We define OTD as fulfilled demand / total demand. Inventory position is measured at the die bank.

Using the rolling horizon simulation framework presented in subsection 3.2, we generate a base demand signal for both LP and 2SP. With this base demand signal, we run the rolling horizon simulation for 26 periods. The simulation is iterated 10 times. The computations are performed on a 12th Gen Intel(R) Core(TM) i5-1250 processor, resulting in approximately 0.5 seconds to calculate one planning epoch for LP, and approximately 0.75 seconds for 2SP. Both LP and 2SP are solved using Gurobi (Gurobi Optimization, LLC 2025).

The plan stability performance of LP and 2SP, including the 95% confidence interval, is depicted in Figure 1. When comparing the stability over time between LP and 2SP, it is evident that 2SP results in a higher average stability score while continuously providing stable stability scores. LP stability performance fluctuates heavily over time and shows a wider confidence interval. The main driver of this fluctuating performance is the nonstationary demand over time, which is caused by the updating of demand in the rolling horizon setting outlined in subsubsection 3.1.1. Recall that demand per product is updated over time using a probability based on historic data. The probability will result in different amounts of demand changes over time. In the event of a relatively large number of changes, LP is likely to make more changes to the plan to remain optimal, resulting in a lower plan stability score.

In Figure 2 we show results for the KPIs OTD and inventory position, including the 95% confidence. For both KPIs, the average performance of all products per period is depicted to facilitate easy comparison of both methods. Both methods show similar OTD performance. The mean of 2SP is slightly higher than the mean of LP, but as the confidence intervals are overlapping, the difference is not statistically significant, and we conclude that OTD performance is similar for both methods. As the first two objective layers in both models strive to maximize the service level, both methods perform equally. The results for inventory position show that LP builds inventory over time, while 2SP stabilizes over time. The mean of 2SP is below the mean of LP. The confidence intervals are not overlapping for 60% of the time periods. Therefore, we conclude that 2SP performs slightly better on inventory position compared to LP. As 2SP has more information on the distribution of the demand uncertainty (through the demand scenarios), it can better anticipate demand changes, resulting in approved base stock levels compared to LP. LP might treat demand changes less conservatively than 2SP, leading to higher inventory levels.

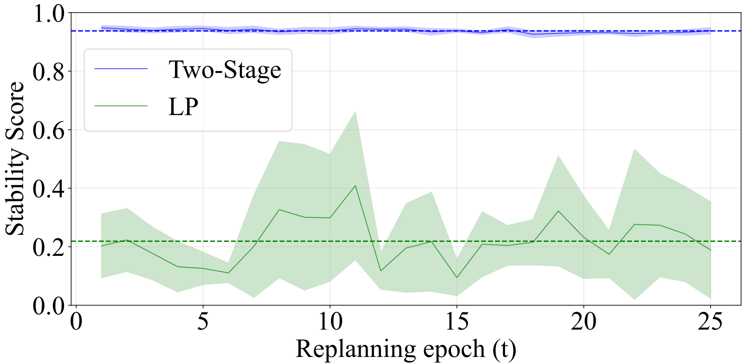
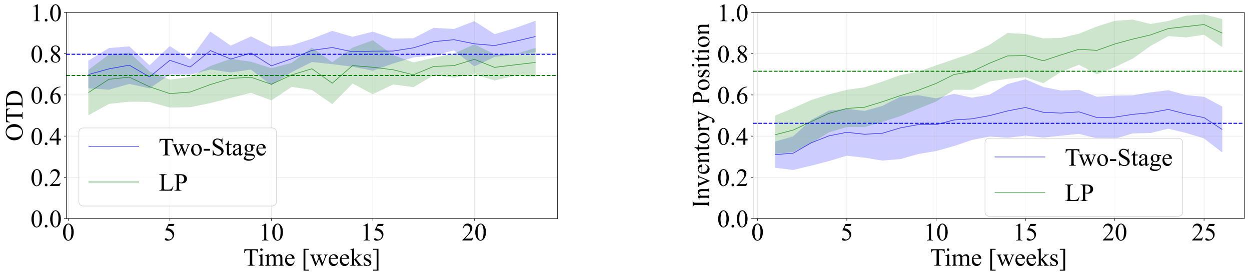


Figure 1: Planning stability including 95% confidence interval per replanning epoch for LP and 2SP.



(a) Average OTD over time.

(b) Average inventory position over time.

Figure 2: OTD and inventory position including 95% confidence intervals over time for LP and 2SP.

6 DISCUSSION

Originality. We propose to use 2SP to increase plan stability in semiconductor supply chain wafer planning. The proposed method is more flexible compared to commonly used methods such as chance-constraint programming. While 2SP shows promising results to improve plan stability, other methods are worth investigating. Especially, if we want to refine the objective from improving plan stability to reacting solely to significant demand changes.

Methodology. We model demand uncertainty by fitting distributions to historic demand data. As such, we can capture the demand pattern in the stochastic demand signal used in our modeling approach. The method can be used for high-demand products. It would be of interest to further improve the method's performance and extend it to ramp-up and ramp-down products. To update demand in the rolling horizon, we use discrete probabilities and normally distributed demand changes. This method could cause overfitting of past demand because of the use of a discrete distribution. Furthermore, if the number of products is increased, it should be investigated whether normally distributed demand changes reflect all products. We apply 2SP to improve plan stability compared to LP. Where only one demand scenario is provided to LP, a set of scenarios is provided to 2SP. As this work is a first exploratory study into using 2SP to improve plan stability, the effect of the number of scenarios and the associated probabilities has not yet been researched. As this is a non-trivial task, this is highly important for future research.

Application. We apply distribution fitting on historic demand data to reflect the demand patterns during demand generation. This stochastic demand signal is used to apply 2SP and compare its performance with

LP based on the KPIs plan stability, OTD, and inventory position. The rolling horizon simulation framework determines the OTD and inventory position performance. We tested the framework using a horizon of 26 weeks and iterated the simulation 10 times. As demand uncertainty increases over time, it would be interesting to explore the performance of the proposed method over a longer horizon. Furthermore, 10 iterations is only a limited number and should be increased in the future to increase the statistical significance of the results. The proposed framework can be used for other use cases in a rolling horizon setting under demand uncertainty. We want to emphasize that these conclusions are based solely on data from NXP Semiconductors N.V. Therefore, when replicating the experiments at other companies, the conclusions must be reevaluated.

7 CONCLUSIONS

In this paper, we propose to use a 2SP method to improve plan stability in a rolling horizon setting compared to LP. To model demand uncertainty, we propose a new method to incorporate demand patterns in the stochastic demand signal through fitting distributions to historic demand data. The method is demonstrated by analyzing the plan generation performance of 2SP and LP using a rolling horizon simulation model of the front-end semiconductor supply chain. The proposed demand generation method, which models demand arrivals using an MC approach and demand sizes using a Gamma distribution, can be used for all demand patterns (intermittent, lumpy, smooth, and erratic). For the first two patterns, distributions for 85% of products in the use case can be fitted. For the other two patterns, a performance of 83% is achieved.

Based on the simulation results, we conclude that 2SP is a promising method to generate less nervous wafer plans compared to LP. While both methods show similar OTD performance, 2SP slightly outperforms LP on inventory position.

For future research, it would be of interest to improve and extend the demand generation method to ramp-up and ramp-down products. The method to update demand in the rolling horizon setting must be validated for a larger set of products. Furthermore, it would be interesting to include supply uncertainty in the model, including cycle time and yield uncertainty. Lastly, the current 2SP implementation should be refined. The number of scenarios presented to 2SP and the associated probabilities should be optimized. Next to improving 2SP, exploring other stochastic optimization methods is an interesting avenue for future research.

ACKNOWLEDGMENTS

The project AIMS5.0 is supported by the Chips Joint Undertaking and its members, including the top-up funding by National Funding Authorities from involved countries under grant agreement no. 101112089.

REFERENCES

De Kok, T., and K. Inderfurth. 1997. “Nervousness in Inventory Management: Comparison of Basic Control Rules”. *European Journal of Operational Research* 103(1):55–82.

Friese, F., and S. Helber. 2023. “A Framework for Multi-Objective Stochastic Lot Sizing With Multiple Decision Stages”. Hannover Economic Papers (HEP) No. 708, Leibniz Universität Hannover, Wirtschaftswissenschaftliche Fakultät, Hannover.

Gurobi Optimization, LLC 2025. “Gurobi Optimizer Reference Manual”. "https://www.gurobi.com", accessed 27.07.2025.

Kabak, K. E., and A. M. Ornek. 2009. “An Improved Metric for Measuring Multi-Item Multi-Level Schedule Instability Under Rolling Schedules”. *Computers & Industrial Engineering* 56(2):691–707.

Kimms, A. 1998. “Stability Measures for Rolling Schedules With Applications to Capacity Expansion Planning, Master Production Scheduling, and Lot Sizing”. *Omega* 26(3):355–366.

Koca, E., H. Yaman, and M. S. Aktürk. 2018. “Stochastic Lot Sizing Problem With Nervousness Considerations”. *Computers & Operations Research* 94:23–37.

Lin, P.-C., and R. Uzsoy. 2016. “Chance-Constrained Formulations in Rolling Horizon Production Planning: an Experimental Study”. *International Journal of Production Research* 54(13):3927–3942.

Mönch, L., R. Uzsoy, and J. W. 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.

Özelkan, E. C., S. Torabzadeh, E. Demirel, and C. Lim. 2023. “Bi-Objective Aggregate Production Planning for Managing Plan Stability”. *Computers & Industrial Engineering* 178:109105.

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

Prak, D., and P. Rogetzer. 2022. “Timing Intermittent Demand With Time-Varying Order-Up-To Levels”. *European Journal of Operational Research* 303(3):1126–1136.

Rashidi, E., T. H. Bhuiyan, and S. J. Mason. 2024. “Production Planning for Semiconductor Manufacturing Under Demand and Yield Uncertainty”. *Computers & Industrial Engineering* 196:110403.

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.

Sáez, P., C. Herrera, and V. Parada. 2023. “Reducing Nervousness in Master Production Planning: a Systematic Approach Incorporating Product-Driven Strategies”. *Algorithms* 16(8):386.

Syntetos, A. A., M. Z. Babai, and N. Altay. 2012. “On the Demand Distributions of Spare Parts”. *International Journal of Production Research* 50(8):2101–2117.

Syntetos, A. A., J. E. Boylan, and J. Croston. 2005. “On the Categorization of Demand Patterns”. *Journal of the operational research society* 56(5):495–503.

Turrini, L., and J. Meissner. 2019. “Spare Parts Inventory Management: New Evidence From Distribution Fitting”. *European Journal of Operational Research* 273(1):118–130.

Ziarnetzky, T., L. Mönch, and R. Uzsoy. 2018. “Rolling Horizon, Multi-Product Production Planning With Chance Constraints and Forecast Evolution for Wafer Fabs”. *International Journal of Production Research* 56(18):6112–6134.

Ziarnetzky, T., L. Mönch, and R. Uzsoy. 2019. “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

ERIC WEIJERS is a doctoral candidate in the Department of Supply Chain Operations at NXP Semiconductors N.V. and in the Department of Industrial Engineering and Innovation Sciences at the Eindhoven University of Technology (TU/e). He obtained his master’s degree in Operations Management & Logistics at TU/e. His primary research interest is in the area of simulation modeling and (stochastic) optimization in supply chain management. His email address is eric.weijers@nxp.com.

NINO SLUIJTER recently obtained his master’s degree in Operations Management & Logistics at the Eindhoven University of Technology, completing his graduation project at NXP Semiconductors N.V. His expertise is in simulation and demand modeling. His email address is ninosluijter@gmail.com.

GIJS HOGERS is a student in the Department of Econometrics and Operations Research at Tilburg University and a master’s thesis intern at NXP Semiconductors N.V. His primary research interest is decision-making under uncertainty in complex supply chain environments. His email address is gijseersel@gmail.com.

KAI SCHELTHOFF is the head of the Supply Chain Innovation team at NXP Semiconductors N.V. He obtained his PhD at the Karlsruhe Institute of Technology about data-driven cycle time estimation in semiconductor wafer fabrication using a concatenated machine learning approach, in collaboration with Robert Bosch GmbH. He is an expert in machine learning applications for estimation, classification, and optimization in Supply Chain Operations and Manufacturing. His email address is kai.schelthoff@nxp.com.

IVO ADAN is a Full Professor in the section Operations, Planning, Accounting and Control (department of Industrial Engineering & Innovation Sciences) at Eindhoven University of Technology (TU/e) and holds the Manufacturing Networks chair. His expertise and tuition areas include probability theory / statistics, operations research, manufacturing networks, stochastic operations research, and queueing models. His e-mail address is i.adan@tue.nl.

WILLEM VAN JAARSVELD is an Associate Professor in Stochastic Optimization and Machine Learning at Eindhoven University of Technology (TU/e). His main research interest is stochastic optimization, using a diverse set of methodologies including Deep Reinforcement Learning, Stochastic Programming, and Dynamic Programming. Application areas include data-driven inventory control, production planning, supply chain management, and maintenance logistics. His email address is w.l.v.jaarsveld@tue.nl.