Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

INCREASING SUPPLY CHAIN ROBUSTNESS DURING ALLOCATION IN A JUST-IN-TIME SUPPLY SET-UP

Volker Dörrsam

Automotive Digitalization & Supply Chain Excellence
Infineon Technologies AG
Am Campeon 1-15
85579 Neubiberg, GERMANY

Jan-Philip Erdmann
Patrick Moder

Supply Chain Engineering Innovation
Infineon Technologies AG
Am Campeon 1-15
85579 Neubiberg, GERMANY

ABSTRACT

In situations of scarcity, that is when demand exceeds available supply, a stable allocation of capacities among customers contributes to a more robust supply chain behavior. Given the input of available capacities in the first place, this paper presents an analytical approach that models its smooth tactical distribution by accounting for product-level deviations. The proposed model may serve as input for allocation determination in situations with demand surges. Simulating the model and conducting experiments using real-world data from a globally acting semiconductor manufacturer, it provides empirical evidence of results in terms of supply chain stability. Still, the proposed model ensures sufficient flexibility due to well-defined target inventory levels.

1 INTRODUCTION

When managing order fulfillment in presence of scarce supply or exceeding demand, suppliers allocate available capacity quotas (available to promise, ATP) to a proportion of customers. Although there exists a broad body of research about frameworks, heuristics and analytical approaches to distribute available supply to customers according to their demand, such approaches prove to be suboptimal in times of demand surges, supply scarcity or supply chain uncertainties in general (Kleindorfer and Saad 2005; Vogel and Meyr 2015; Kloos and Pibernik 2020; Niranjan et al. 2022). While suppliers may seek to maximize expected profit, customers are interested in receiving orders on time in full (OTIF), thus considering any delay or supply reduction as deviation from their – personal – optimum. Customers try to mitigate these disruption risks by maintaining sourcing strategies (Jain et al. 2022), nevertheless tending to irrational behavior, i.e., shortage gaming and order inflation, in situations of scarce supply with the aim to receive as many orders as possible OTIF (Sterman and Dogan 2015; Armony and Plambeck 2005). To counteract, suppliers usually establish customer segmentation according to priorities and customer life time value, among other criteria. In addition, suppliers may mitigate the negative impact of exaggerated orders by correcting their projected demand according to an expected *true* demand or schedule padding (Niranjan et al. 2022). This way, maximized profit is targeted while satisfying the customers that matter most to the supplier on hand. This strategy is commonly supported by advanced planning systems (APS) in a sequential manner with increasing granularity, which constitutes advanced ATP (Pibernik 2005). To ensure flexibility, in some circumstances it is adjusted by subject matter experts according to tacit knowledge and individual preferences. When customer demand surpasses available capacity, affected products are put ”on allocation”, thus the supplier closely monitors the available capacity split to customers that request just-in-time (JIT) delivery. This way, available capacity peaks (and drops) immediately influence the share of supply per customer. Nevertheless, given a longer observation period, this procedure has proven adverse with regard

to supply chain stability (Spiliotopoulou et al. 2022; Vogel and Meyr 2015), i.e., potentially evoking the bullwhip effect (see (Sterman 1989) for a behavioral perspective or (Lee et al. 1997) for an operational viewpoint). Hence, this study builds on the argument that, when demand exceeds available supply, a *stable* distribution of ATP among customers would contribute to a more robust supply chain behavior and increased service levels in the long run. Given the situation of rule-based and JIT-driven ATP in the first place, this paper presents an analytical approach that models the smooth tactical distribution of such ATP by accounting for product-level deviations. In turn, the proposed model may serve as input for allocation determination in situations with demand surges. Simulating the model and conducting experiments using real-world data from a globally acting semiconductor manufacturer, it provides empirical evidence of results in terms of supply chain stability. Still, the proposed model ensures sufficient flexibility due to well-defined target inventory levels.

The remainder of this paper is structured as follows. Semiconductor supply chain planning as research environment and related literature streams covering allocation optimization are presented in Section 2. After describing the design of experiment in Section 3, the model developed in the study on hand is introduced in Section 4. The results of testing this model in an empirical setting are presented in Section 5. A thorough discussion of these results is provided in Section 6, before concluding with limitations and future research avenues in Section 7.

2 RELATED RESEARCH

In semiconductor supply chain, scarce capacity and demand surges are common problem statements. Besides being extensively described in popular media as "Chip shortage" since the Covid-19 outbreak recently (Yang and Sohn 2021), such situations have already been addressed in the past on a regular basis. In addition to external factors, i.e., pandemics, natural disasters or other disruption sources (Kleindorfer and Saad 2005; Simchi-Levi et al. 2014), complexity related to the inherent characteristics of semiconductor manufacturing (and supply chain) is causing such uncertainties (Chien and Zheng 2012; Mousavi et al. 2019). Characteristics include long cycle times, global manufacturing networks and capital-intense production facilities. In addition, the upstream position of semiconductor manufacturers increases impacts of the bullwhip effect that accounts for additional uncertainties in order fulfillment (Lee et al. 1997). Semiconductor manufacturers handle such situations with advanced planning systems, safety stocks and contractual commitments between their customers, usually applying control decisions that span the entire organization with detailed granularity (Kempf 2004; Fordyce et al. 2011). Nevertheless, in situations where demand exceeds supply, some sort of triage is inevitable. In these situations of tightness, supply is allocated to customers to optimally satisfy their orders and mitigate negative impacts on customers and the customers' customers in turn. Like in other complex supply chain environments, ATP systems integrated with Enterprise Resource Planning systems are used commonly in the semiconductor industry in order to describe expected supply on product level and to ensure supply chain flexibility (Seitz and Grunow 2017). Supply network planning serves as input for demand fulfillment, where in the first place forecasts from customers are balanced against available ATP during allocation planning. That result from allocation planning in turn serves as an input to balance this so-called *allocated ATP* (Kilger and Meyr 2008) against actually incoming orders by customers. The latter second sub-process of demand fulfillment is often defined as order promising (cf. (Seitz et al. 2016) (Seitz and Grunow 2017) or (Seitz et al. 2020) for more details). The approach introduced in this paper aims at a more stable allocated ATP as input for order promising, which is achieved through an automated smoothening of ATP on product granularity (as shown in this paper) before the manual allocation of ATP on customer granularity proceeds (as proposed in the outlook). Besides capacity allocation approaches (Mallik and Harker 2004; Ng et al. 2010; Chien et al. 2013; Mönch et al. 2020), several optimization approaches for allocating ATP to customers are recently studied in the semiconductor supply chain context, the most relevant of them being summarized in the following.

While focusing on allocating production lots to orders, (Ng et al. 2010) also reflect on the impacts of over- and underfulfillment of orders on hand. In their study, penalties for unfulfilled or overfulfilled demand

are introduced for the mixed-integer optimization problem that is solved via branch-and-price and Benders decomposition. Here, overfulfillment results in opportunity costs for exceeding demanded quantities, while unfulfillment results in lost sales, customer dissatisfaction and a higher probability of customer churn. (Mousavi et al. 2019) developed an optimization approach for product allocation to customers in the semiconductor supply chain in the form of a Mixed Integer Linear Program. The model is optimizing the allocation of ATP and buffer stock with different priorities on seller level. The model has a multi-objective character with the objectives to maximize the allocated quantities from buffer and ATP on the one hand and to ensure a remaining buffer on the other hand. The weights of these functions are varying to analyze their impact on the allocation of products to customers. As a verification, the model's output is compared to the allocation, which is created by Supply Chain Planners (SCPs). Since the optimization is not implemented as a recommendation tool for SCPs, an implementation is described as future work. (Mousavi et al. 2019) conclude that the mathematical description of the customer allocation process, which is developed in the paper, are close to the SCPs' allocation and follow the general allocation logic.

(Ziarnetzky et al. 2019) simulate the optimized placement of engineering lots within the semiconductor supply chain. Engineering lots are reserved quantities of semiconductors for research and development purposes. Therefore, the lots reduce the quantity which can be allocated to customers. This case can be seen as one of the different experiment scenarios conducted in this paper. The authors' optimization is aiming for a reduced impact on the customer-focused production of semiconductors by minimizing the total costs and analyzing the impact of engineering costs. The model is based on the general semiconductor supply chain cost optimization, which was developed by (Ziarnetzky and Mönch 2016). It adds the engineering costs in different engineering lot configurations and compares different prioritization settings for engineering lots.

Within the literature stream of the semiconductor supply chain, further optimization approaches exist that are related but follow a different approach than this paper. (Framinan and Perez-Gonzalez 2016) analyse ATP systems in the industry within a high customer heterogenity environment. They describe the different modes and configurations in ATP planning systems: different input granularities (plant vs. customer level), different planning rules (no allocation vs. customer allocation), different ATP consumption modes and the optional use of ATP reallocation (Framinan and Perez-Gonzalez 2016). On another granularity which is the capacity allocation of machines and tools to products in the photolithography area, (Ghasemi et al. 2018) use a genetic algorithm to optimize this allocation problem to maximize the loading level of machines. They conclude that the algorithm is efficient and potentially helpful if stochasticity is tackled (Ghasemi et al. 2018). Yet another granularity is analysed by (Deenen et al. 2019). They look at wafer-to-order allocation, in a bin-covering problem, using different algorithms to solve it. Out of the different algorithms, the integer linear program achieves a reduction to 43.736% of the overallocation within the manual allocation which outperforms other analyzed heuristics (Deenen et al. 2019).

Besides semiconductor specific research, there is also a broad body of literature on order allocation and supply chain management under uncertainty in general (see (Govindan et al. 2017) for an extensive review on supply chain network design under uncertainty). While most allocation studies particularly aim for optimizing profit, some also provide ideas around maximizing service quality (see (Pibernik and Yadav 2008) for make-to-order and (Pibernik and Yadav 2009) for make-to-stock settings or (Kloos et al. 2018) for an environment with sales hierarchies. In essence, different dimensions of the allocation and order promising problem are examined, namely coordination of supply chain triads (Hsieh and Wu 2008), information exchange in hierarchical settings (Fleischmann et al. 2020), contracts dependent on service level (Kloos and Pibernik 2020), among others.

To summarize, allocation and order promising are thoroughly discussed in literature, yet with some limitations that the paper on hand aims to overcome. Namely, the presented allocation model addresses both the aspect of supply chain stability and the maintenance of service level quality when supply is scarce. It allows for a flexible supply chain behavior through the precise control of buffer inventory levels, thus avoiding overfulfillment of demand. Still, by utilizing available capacity, the model enables exhaustive

demand fulfillment. The model's decision support capabilities are proven in a semiconductor supply chain setting with real-world data.

3 METHOD

3.1 Research Environment

During allocation phases, existing buffers are depleted and established procedures to determine target buffer levels are no longer applicable. In parallel, increasing variability in both demand and supply occurs, adding instability to supply chain planning. On the demand side, customers try to balance shortages from different suppliers and prefer to shift production from high volumes to the most profitable (low volume) variants. These shifts inherently vary over time as single components may determine the weekly availability of subcomponents. In parallel, customers may place higher forecast or orders to hedge against uncertainty, oftentimes beyond the levels of the actual demand. On the supplier side, i.e., the semiconductor manufacturer in the study on hand, challenges arise due to high manufacturing leadtime. Such long leadtimes add another layer of complexity when determining what ratio of the order level per product can be assumed as *true* demand: decisions about production strategies and product mix at wafer start will impact available supply in periods often more than six months ahead.

Usually, target levels of buffers are calculated by *reaches of demand*, i.e., multiple of weeks demand. With hoarding, phantom ordering (Sterman and Dogan 2015) or duplicate ordering (Armony and Plambeck 2005) behavior by customers during allocation, demand is overstated, which likewise holds for the buffer reach. Hence, a strategy that includes buffer building priority cannot be kept with overall supply being insufficient. This results in the depletion of buffers on finished and semi-finished goods level. However, as soon as internal buffers get depleted, one can observe an increase in output variability.

Again, due to long production times in semiconductor manufacturing, production lots have a high chance to be imposed by multiple shortages, priority changes or other deviations that most probably lead to altering the initial plan. Multiple kinds of deviations already exist in *normal* situations, yet there is usually little impact on the inputs for an ATP engine due to the maintenance of many buffers. However, a broader portfolio of products is affected when facing unexpected deviations, such as the COVID-19 outbreak and natural disasters that lead to production shutdowns, transportation limitations or customs restrictions (cf. (Kleindorfer and Saad 2005)). Therefore, any plan that does not account for existing or projected variabilities is subject to high error. Human intervention during allocation is one strategy to overcome the limitation of the systems' inability to automatically adapt in a flexible manner. Human allocation experts use two procedures to stabilize commitments towards customers:

1. Using runrates instead of treating the planning output as is. The main benefit of using runrates is that changes in supply are dampened by some averaging effect. An average of non-negative numbers is only possible if some portion of the supply drops.
2. Building buffer to react on unplanned supply shortages. However, buffer building beyond the absolute minimum inherently tightens the supply.

Therefore, a high frequency of reviewing and adjusting to meet an optimal stabilized allocation is required for both runrate levels and buffers. Nevertheless, skilled allocation experts are a scarce resource, thus the need for automation is high. The algorithms proposed in this study relief the reviewing and adjustment processes. Instead of using the entire ATP for allocation per period, it is modified by an optimization model that introduces stable ATP quantity levels, which are monotonously growing if possible. This leads to the building of buffer quantities which are limited using a minimum buffer definition in the model. The minimum buffer increases with the model timeline to address the increasing uncertainty of input values with the time horizon. However, in the early weeks, the model ensures that available quantities can be shipped by using a rolling target buffer ramp, which serves as a *virtual* buffer for uncertain future periods. Hence, we ensure a more stable allocation in the long run and allow allocation experts to focus on balancing

other areas, e.g., adjusting the product mix. Moreover, within the described environment, designing robust supply chains is essential to ensure sustainable supply to customers. The ATP-driven allocation situation is dependent on supply stability to ensure stable runrates. In this paper, an approach to increase stability is presented. Due to semiconductor shortages, critical supply chains currently tend to move from a JIT towards a Just-in-Case (JIC) approach. Thus, confirmed quantities can be met with a higher reliability since short term reductions and supply losses can be compensated. This paper recommends an approach using a modified ATP as a JIC setting. Instead of using the entire ATP for allocation per period, it is modified by an optimization model by introducing stable ATP quantity levels that are monotonously growing if possible. Hence, we ensure a more stable allocation in the long run.

3.2 Experimental Design

This paper describes the development of such a model as linear optimizer as well as its evaluation in different allocation situations and circumstances. Within the demand fulfillment process, this model is located just before the customer allocation process. As this process is undergoing a transition towards a semi-automated process, the experiment on hand is intended to evaluate the modified ATP in different situations in terms of its stability as input for order promising purposes. Therefore, the conditions of several inputs parameters to the optimization model are varied, as listed in the following.

- The optimization is running on *product* level. All products, which run in the allocation automation, can be taken. Different representative examples were picked for the experiment.
- The *minimum buffer* has two levels and a ramp in between within the investigated time period. The levels and ramp can be designed using the parameters *start week buffer*, *end week buffer* and *supply reach*.
- The *optimization period* is defined by the parameters *freeze end* and *opt end*.
- *Inventory Weight* affects the model's prioritization to build inventory. It is used in the target function of the model.
- *Discount* describes the importance of each period's contribution to the target function. The higher the discount, the less important are later periods in the optimization model

Here, the minimum buffer serves to cope with increasing uncertainty over the optimization time frame. It is stated per relative week from the starting week, i.e., relative week 0. Typically, in the beginning, zero buffer is used to make as much supply available to fulfil customer needs. This period is followed by a (gradual) ramp of the minimum buffer up to a predefined target level. Finally, the minimum buffer levels off to a predefined target level. When to start and when to end the ramp and what level of target buffer to use, is negotiated between the enterprise functions involved in allocation decision making. As time passes and another optimization run is performed, the minimum buffer for a specific calendar week is then automatically reduced as target figures are bound to relative weeks.

The experiments are conducted using the `pywraplp` solver of Google OR Tools in *Python* with 479 selected products from the real-world environment over a period of 13 weeks.

4 MODEL

The idea of the optimization model is the introduction of constant *supply levels* sl_t, which are only allowed to be increased within the *optimization period* T under normal circumstances. These circumstances are defined by the normal allocation situation in which the cumulated ATP_t does not exceed the cumulated *Requested Material Arrival Date Quantity* $RMAD_t$. In these cases, the model is allowed to reduce the *supply level* by changing the direction, which is done in Definition (1) and Constraint (6). The *supply levels* are maximized while the inventory in minimized in the Target Function (5). It uses *Net Present Values* npv_t defined in (4) to prioritize early weeks and an *inventory weight* $w^{inventory}$ defined in (3) as parameters. The decision variables of the model are sl_t and *Inventory* i_t. The inventory before the first period is a boundary

condition and requires to be non-negative (2). It is calculated as a minimum of ATP, *RMAD* and original target allocation (TA) from the week before the optimization period. The growth pattern of the *supply levels* is realized in Constraints (6) and (7). The Inventory Constraint (8) calculates the new inventory from the old inventory and in- and output. *Target Inventory Buffer* i_t^{target} can be defined as desired and is then used in Constraint (9) as a minimum for the actual inventory in the model. To not have supply that exceeds orders in the model, Constraint (10) defines the cumulated *RMAD* as maximum bound for cumulated *supply levels*. Constraint (11) defines the limits of the *supply levels* sl_t to 0 as lower bound and infinity as upper bound by default. The optimization period T are the weeks in which the optimization is performed, i.e., 13 weeks in the study on hand.

Definitions

$$direction_t = \begin{cases} -1, & \text{if} \sum_{\tau=0}^{t} RMAD_\tau + i_\tau^{target} - \sum_{\tau=0}^{t} ATP_\tau < 0 \\ 1, & otherwise \end{cases} \quad \forall t, \tau \in T \tag{1}$$

$$inventory_0 \geq 0 \tag{2}$$

$$w^{inventory} \geq 0 \tag{3}$$

$$npv_t = npv_{t-1} * (1 - discount) \quad \forall t \in T \tag{4}$$

Target Function

$$\max \sum_t npv_t * (sl_t - w^{inventory} * i_t) \tag{5}$$

s.t.

$$\Delta sl_t = (sl_t - sl_{t-1}) * direction_t \quad \forall t \in T \tag{6}$$

$$\Delta sl_t \geq 0 \quad \forall t \in T \tag{7}$$

$$i_t = i_{t-1} + ATP_t - sl_t \quad \forall t \in T \tag{8}$$

$$i_t \geq i_t^{target} \geq 0 \quad \forall t \in T \tag{9}$$

$$\sum_{t=0}^{\tau} sl_t \leq \sum_{t=0}^{\tau} RMAD_t \quad \forall t, \tau \in T \tag{10}$$

$$sl^{lower} \leq sl_t \leq sl^{upper} \quad \forall t \in T \tag{11}$$

5 RESULTS

The experiment is analysed in four stages that focus on different aspects with increasing granularity:

1. Mean and variation comparison of all products and time shifts
2. Impact of buffer on the stability between consecutive time shifts
3. Comparison with the original allocation in a split between overallocated and non-overallocated products
4. Comparison of simulated deliveries with the supply and manual allocation

The analysis is based on real data of product and region combinations which is anonymized for this purpose. Within the first three stages, the analysis focuses on a rolling time frame. In a first iteration, weeks 0-13, i.e., a quarter, is analyzed, while in all upcoming 13 iterations the observed time window gets shifted to the future by one week per iteration. Hence, the overall observation period is half a year, i.e., 26 weeks. Incoming supply is optimized using the presented model both with and without a buffer. Please note that *optimized supply* refers to using the model presented in this paper; this holds for all four stages of the experiment. The results are then compared with original ATP and manual allocation. In the last stage, actual deliveries are simulated and their variability is analysed.

First, the mean values of the 13 week time frames and their coefficient of variation are compared for optimized supply both with weekly and without buffer, manual allocation and original ATP. This way, the stability of the optimization is examined. In Figure 1, these four outcomes are visualized. It is found that the mean results are similar except for the manual allocation, which has 10.2% higher values than the others' mean. The coefficient of variation is relatively low for the optimized *supply levels* compared to the manual allocation and the original ATP.

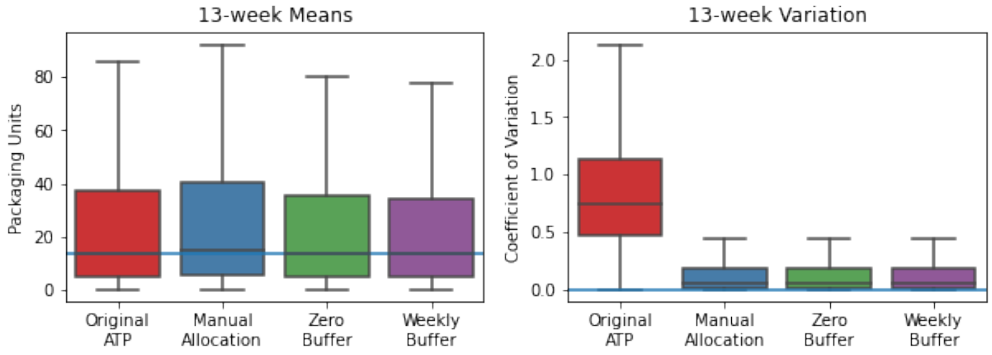

Figure 1: Comparison of mean and coefficient of variation between optimized *supply levels*, ATP and manual allocation.

Second, the differences of time shifts relative to the total mean of the weekly *supply level* are compared and analysed regarding their stability over time as shown in Figure 2. The analysis is focused on the weeks after the freeze fence of four weeks to gain a more detailed view on the mid-term stability. The freeze fence is further analysed in the discussion. Original ATP shows a high fluctuation over time with high difference percentages without a visible trend of increase of decrease of relative difference, which is visible in Figure 2. Manual allocation has a lower variation with the majority between -1% and +1% per week. The median of each time shift is 0 in every week. The results for the optimized *supply levels* start with a higher variation in the first week with a trend towards a lower fluctuation over time. Compared to the manual allocation, it is higher by a small margin at the beginning of the analysed time and similar and even lower for later weeks in the time period. The medians are above zero for both the optimization with and without a weekly buffer.

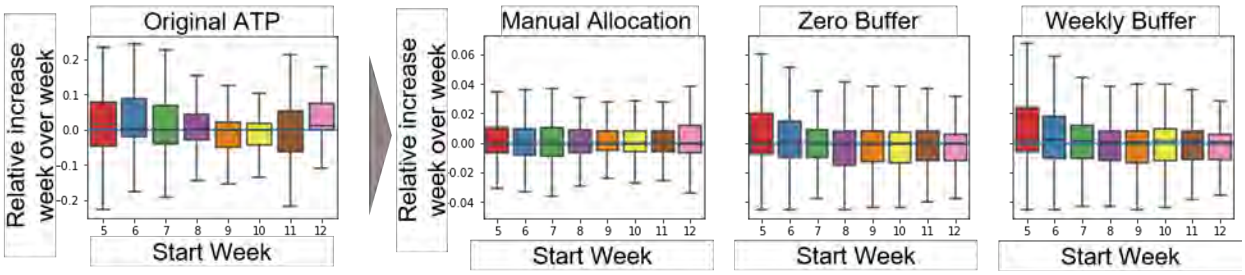

Figure 2: Time shift variability of original ATP compared with optimized *supply levels* and manual allocation.

Third, the data set was split into cases with overallocation and cases without overallocation in the manual allocation setting to have a more detailed look into the performance comparison of the optimized results with the manual ones, which is visualized in Figure 3. In the comparison of means for non-overallocated products, a generally increased mean for the optimized products scenarios can be detected compared to the manual allocation setting. For overallocated products, the mean of the manual setting is higher than the

ones of the optimized *supply level* allocation, i.e., more packaging units are overallocated in the manual setting.

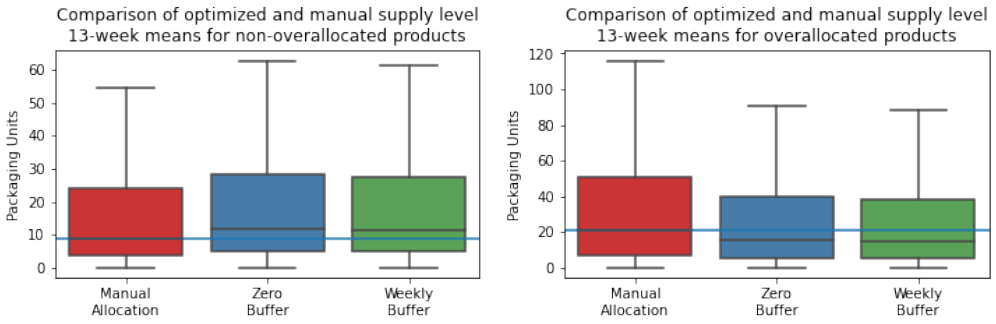

Figure 3: Overallocation analysis and average *supply level* comparison for overallocated and non-overallocated products.

Fourth, the actual supply based on the optimized cases was simulated to see the behavior of *supply levels* with progressing time. The simulation for the ATP is shown in Figure 4 as a comparison. This is again compared to manual allocation. The visualization shows – similar to stage 2 – relative fluctuations between each of the weeks. Here, the optimized results have a lower fluctuation of deliveries compared to the original ATP and the manual allocation. Their median is 0 for all weeks. Comparing the optimized supplies, within weeks 5-8, some cases with a positive delta occur in the buffered setting compared to the case without buffer-building. Fluctuations of original ATP are approximately ten times higher compared to the manual allocation. Both do not have a trend like the optimized ones. The manual allocation has a high amount or variety in fluctuation for week 6.

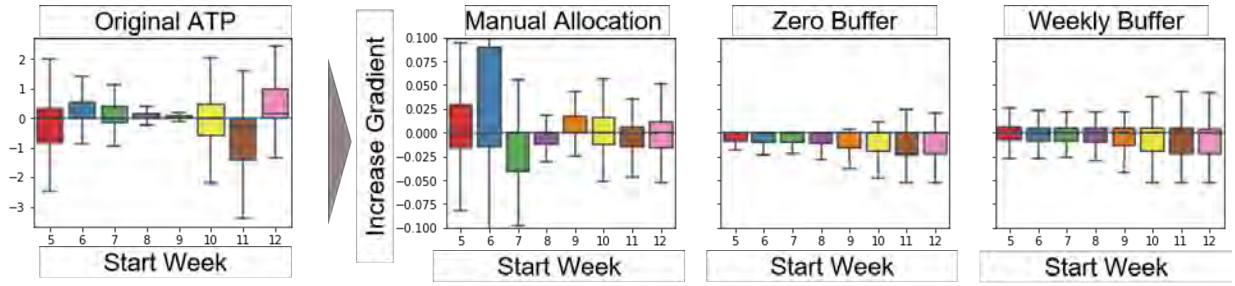

Figure 4: Analysis of simulated supply within the first 13 weeks of optimization used after freeze phase for the original ATP and the simulated supply for optimized and manual scenarios.

6 DISCUSSION

The discussion follows the four stages introduced above. Stage 1 compares the means and the coefficients of variation of the rolling 13-week time series for the four cases optimized *supply level* with buffer building, optimized *supply level* without buffer building, manual allocation and original ATP. The mean of the weekly buffer optimization is in general a bit lower than the zero-buffer cases due to the quantity which is put into inventory to absorb unexpected changes. The mean of the original ATP is similar to the zero-buffer case, which shows that most of the quantity is used. Thus, the optimized supply configuration does not unnecessarily hold back quantity or decreasing sales and profits. The manual allocation is higher than the other cases, since the allocation for some cases is higher than the actual ATP. This can have several reasons, which is further discussed in Stage 4. The original aim of the ATP optimization is an introduction

of more stable and mostly monotonously increasing *supply levels*. Therefore, the variation within the 13-week rolling time series should be lower than both, the manual allocation and the original ATP, which is proven within the analysis on hand. With a decreased variation, the supply is more stable in general, which not only increases customer satisfaction but also decreases the impact of unforeseen supply changes on delivery reliability in the long run.

Stage 2 focuses further on *supply level* variability by examining relative increases and decreases between the different 13-week time series with increasing starting weeks. The diagrams show high fluctuation of the ramp up phase in weeks 0-4 for the optimized results (not shown due to scaling reasons), which appears due to pre-optimization (inventory calculation) and hence not problematic for applications in reality. After this phase, variability is still higher than in the manual allocation but it shows a trend of further reduction. This appears due to necessary adaptions towards starting inventory after the freeze fence, which reduce with time for the weekly buffer and the zero buffer case. The manual allocation has a lower variability at the beginning but it does not show a trend of variability reduction with time. All of the three methods of *supply level* placement show a lower variability than the original ATP. The high ATP variability results from fluctuations: for instance, the supply can be low in some weeks due to engineering lots or production shifts to other products, while peaking in other weeks leading to a high variability. This effect is the original motivation for the introduction of the optimization model. Due to the reduced variability, the optimized supply is more robust than the original one. Short term supply problems only affect the inventory and not the *supply level* itself. Therefore, confirmations are not affected and the delivery plan can be fulfilled in cases of supply changes.

In Stage 3, overallocation was analysed to explain the reason for a higher manual allocation compared to the optimized results from Stage 1. Therefore, the product data was first split into a data set with overallocated products and one without overallocation to analyse them separately. First, the analysis of the manual allocation with overallocation shows an increase of overallocation over time. This is since manual allocation is not reacting fast enough to adapt to supply changes accordingly, which in turn leads to overallocation. Then, the overallocation case shows a higher mean of manual allocation *supply levels* than with the optimized *supply level* configuration. This additional allocated quantity cannot be fulfilled with the current supply picture, which makes the optimized configuration the preferable option, although being lower. Furthermore, this explains the higher *supply level* mean for the manual allocation in Stage 1. For the data set without overallocation, the manual allocation is lower than the optimized one in general. This shows that the supply is not used optimally and unnecessary buffer is built up. Therefore, the optimized *supply level* build-up outperforms manual allocation in every case.

In Stage 4, a simulation of the supply is developed to investigate operating the *supply level* optimization. Here, weekly variation is compared to the actual manual allocation and original ATP without using 13-week time series. In this analysis, it becomes visible that – for the weekly buffer case and the zero buffer case – variability of the optimized *supply level* is significantly lower than for manual allocation and original ATP. This proves that the general aim of achieving more stable *supply levels* is working. Furthermore, the effect of buffer building is clearly visible. With time, the target buffer for each week decreases since it gets closer to the freeze fence. Therefore, buffer inventory can be released for allocation, which explains the higher *supply level* gradients shown in Figure 4.

In general, the four stage analysis shows the applicability of an optimized supply build-up. It fulfills both goals, (i) a decrease of variability to increase robustness towards unexpected supply changes, and (ii) an optimized allocation level to increase supply quantities (and thus service level) to the customer.

Concerning a potential implementation of the optimizer into the Infineon Supply Chain, there are several possibilities. On the one hand, the optimization can serve as a recommendation to the SCPs by indicating a maximum usable ATP. However, this option would not decrease the manual effort and only positively affect the supply chain if seriously considered by the SCP. On the other hand, our proposal can be implemented as a replacement of the actual ATP after the planning system and for the following processes as shown in Figure 5. This ensures an automated and stabilized customer allocation with reduced manual effort.

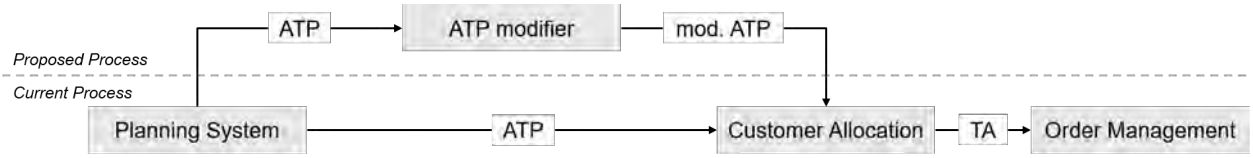

Figure 5: Conceptual proposal of the implementation of modified and stabilized ATP optimization.

7 CONCLUSION

In times of semiconductor shortage, having a robust supply to changes is crucial to ensure delivery reliability. The approach presented in this paper uses an optimization to introduce stable *supply levels* that are derived from the demand and the actual ATP. These stable *supply levels* ensure a higher robustness towards short-term supply losses or shifts and improve customer allocation by reducing quantity fluctuations. To analyze the output of the optimization in form of the *supply levels*, a four-stage result analysis was conducted comparing buffer-building optimization, zero-buffer-optimization, manual allocation and original ATP. The fluctuations of supply are reduced significantly compared to the original ATP and they are similar to the manual allocation. The supply quantity is optimized and therefore avoiding overallocation like it occurs in the manual setting. A supply-simulation, which is shown in Stage 4 shows that the fluctuations of the actual supply are lower than the manual allocation's. This shows that the optimized solutions is suiting the allocation problem on the reviewed granularity. It improves robustness by eliminating overallocation, stabilizing supply and maximizing allocation quantity. There are limitations to the optimization, which need to be respected in a potential application. First, the optimization and the analysis assume that the ATP, which is taken as the model input, matches the actual supply. Forecast inaccuracies might lead to differences between ATP and supply, especially in the long-term view. The benefits of the optimization are in the short-term allocation which has a more accurate ATP compared to long-term forecasts. Second, the optimization is carried-out on product-plant granularity which does not imply relations between distribution centers on the same product. An approach which accumulates the same products in different centers can lead to improvement. This approach was not taken to match the optimization's environment of manual and automated allocation. One solution to that would be an additional pre-processing step which accumulates the quantities of different regions before the optimization which would need to be modified to suit the problem changes. This can be conducted by further research about ATP optimization.

There are different ways of implementing the optimization in the allocation process of a semiconductor supply chain. The recommended one is visualized in Figure 5. One important aspect that needs to be taken into consideration is supply chain planners' reaction on the optimized supply. A less fluctuating supply can only be realized with building and using inventory in different supply phases. Therefore, supply quantities need to be held back from allocation, which reduced the planners' freedom of decision. Therefore, a potential application could focus more on data-driven decision support that respects the boundary conditions of human-technology interaction. This way, it might present an optimized solution to the planner, while keeping the freedom of decision, especially in situations when human cognitive abilities are clearly superior. The environments' reaction to the implementation of the optimization would need to be analyzed within further research to use the optimization's potential to increase supply chain robustness.

REFERENCES

Armony, M., and E. L. Plambeck. 2005. "The Impact of Duplicate Orders on Demand Estimation and Capacity Investment". *Management Science* 51(10):1505–1518.

Chien, C.-F., J.-Z. Wu, and C.-C. Wu. 2013. "A Two-Stage Stochastic Programming Approach for new Tape-Out Allocation Decisions for Demand Fulfillment Planning in Semiconductor Manufacturing". *Flexible Services and Manufacturing Journal* 25(3):286–309.

Chien, C.-F., and J.-N. Zheng. 2012. "Mini–Max Regret Strategy for Robust Capacity Expansion Decisions in Semiconductor Manufacturing". *Journal of Intelligent Manufacturing* 23(6):2151–2159.

Deenen, P. C., J. Adan, J. Stokkermans, I. J. Adan, and A. Akcay. 2019. "Wafer-to-Order Allocation in Semiconductor Back-End Production". In *Proceedings of the 2019 Winter Simulation Conference*, edited by N. Mustafee, K.-H. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, 2360–2371. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Fleischmann, M., K. Kloos, M. Nouri, and R. Pibernik. 2020. "Single-Period Stochastic Demand Fulfillment in Customer Hierarchies". *European Journal of Operational Research* 286(1):250–266.

Fordyce, K., C.-T. Wang, C.-H. Chang, A. Degbotse, B. Denton, P. Lyon, R. J. Milne, R. Orzell, R. Rice, and J. Waite. 2011. "The ongoing Challenge: Creating an Enterprise-Wide Detailed Supply Chain Plan for Semiconductor and Package Operations". In *Planning Production and Inventories in the Extended Enterprise*, edited by K. G. Kempf, P. Keskinocak, and R. Uzsoy, 313–387. New York, New York: Springer.

Framinan, J. M., and P. Perez-Gonzalez. 2016. "Available-To-Promise Systems in the Semiconductor Industry: A Review of Contributions and a Preliminary Experiment". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 2652–2663. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ghasemi, A., C. Heavey, and K. Kabak. 2018, 12. "Implementing a New Genetic Algorithm to Solve the Capacity Allocation Problem in the Photolithography Area". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 3696–3707. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Govindan, K., M. Fattahi, and E. Keyvanshokooh. 2017. "Supply Chain Network Design under Uncertainty: A Comprehensive Review and Future Research Directions". *European Journal of Operational Research* 263(1):108–141.

Hsieh, C.-C., and C.-H. Wu. 2008. "Capacity Allocation, Ordering, and Pricing Decisions in a Supply Chain with Demand and Supply Uncertainties". *European Journal of Operational Research* 184(2):667–684.

Jain, N., K. Girotra, and S. Netessine. 2022. "Recovering Global Supply Chains from Sourcing Interruptions: The Role of Sourcing Strategy". *Manufacturing & Service Operations Management* 24(2):846–863.

Kempf, K. G. 2004. "Control-Oriented Approaches to Supply Chain Management in Semiconductor Manufacturing". In *Proceedings of the 2004 American Control Conference*. June 30th - July 2nd, Boston, Massachusetts, 4563–4576.

Kilger, C., and H. Meyr. 2008. "Demand Fulfilment and ATP". In *Supply Chain Management and Advanced Planning*, edited by H. Stadtler and C. Kilger, 181–198. Heidelberg: Springer.

Kleindorfer, P. R., and G. H. Saad. 2005. "Managing Disruption Risks in Supply Chains". *Production and Operations Management* 14(1):53–68.

Kloos, K., and R. Pibernik. 2020. "Allocation Planning under Service-Level Contracts". *European Journal of Operational Research* 280(1):203–218.

Kloos, K., R. Pibernik, and B. Schulte. 2018. "Allocation Planning in Sales Hierarchies with Stochastic Demand and Service-Level Targets". *Operations Research Spectrum* 41(4):981–1024.

Lee, H. L., V. Padmanabhan, and S. Whang. 1997. "Information Distortion in a Supply Chain: The Bullwhip Effect". *Management Science* 43(4):546–558.

Mallik, S., and P. T. Harker. 2004. "Coordinating Supply Chains with Competition: Capacity Allocation in Semiconductor Manufacturing". *European Journal of Operational Research* 159(2):330–347.

Mönch, L., L. Shen, and J. W. Fowler. 2020. "Heuristics for Order-Lot Pegging in Multi-Fab Settings". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 1742–1752. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Mousavi, B. A., R. Azzouz, and C. Heavey. 2019. "Mathematical Modelling of Products Allocation to Customers for Semiconductor Supply Chain". *Procedia Manufacturing* 38:1042–1049.

Mousavi, B. A., R. Azzouz, C. Heavey, and H. Ehm. 2019. "Simulation-Based Analysis of the Nervousness within Semiconductors Supply Chain Planning: Insight from a Case Study". In *Proceedings of the 2019 Winter Simulation Conference*, edited by N. Mustafee, K.-H. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, 2396–2407. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ng, T. S., Y. Sun, and J. Fowler. 2010. "Semiconductor Lot Allocation using Robust Optimization". *European Journal of Operational Research* 205(3):557–570.

Niranjan, T. T., N. K. Ghosalya, and S. Gavirneni. 2022. "Crying Wolf and a Knowing Wink: A Behavioral Study of Order Inflation and Discounting in Supply Chains". *Production and Operations Management* 31(3):1071–1088.

Pibernik, R. 2005. "Advanced Available-to-Promise: Classification, Selected Methods and Requirements for Operations and Inventory Management". *International Journal of Production Economics* 93-94(1):239–252.

Pibernik, R., and P. Yadav. 2008. "Dynamic Capacity Reservation and Due Date Quoting in a Make-to-Order System". *Naval Research Logistics* 55(7):593–611.

Pibernik, R., and P. Yadav. 2009. "Inventory Reservation and Real-Time Order Promising in a Make-to-Stock System". In *Supply Chain Planning*, edited by H. Meyr and H.-O. Günther, 169–195. Heidelberg: Springer.

Seitz, A., H. Ehm, R. Akkerman, and S. Osman. 2016. "A Robust Supply Chain Planning Framework for Revenue Management in the Semiconductor Industry". *Journal of Revenue and Pricing Management* 15(6):523–533.

Seitz, A., and M. Grunow. 2017. "Increasing Accuracy and Robustness of Order Promises". *International Journal of Production Research* 55(3):656–670.

Seitz, A., M. Grunow, and R. Akkerman. 2020. "Data Driven Supply Allocation to Individual Customers Considering Forecast Bias". *International Journal of Production Economics* 227:107683.

Simchi-Levi, D., W. Schmidt, and Y. Wei. 2014. "From Superstorms to Factory Fires: Managing Unpredictable Supply Chain Disruptions". *Harvard Business Review* 92(1-2):96–101.

Spiliotopoulou, E., K. Donohue, and M. Ç. Gürbüz. 2022. "Ordering Behavior and the Impact of Allocation Mechanisms in an integrated Distribution System". *Production and Operations Management* 31(2):422–441.

Sterman, J. D. 1989. "Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment". *Management Science* 35(3):321–339.

Sterman, J. D., and G. Dogan. 2015. ""I'm not Hoarding, I'm just Stocking Up Before the Hoarders Get Here.": Behavioral Causes of Phantom Ordering in Supply Chains". *Journal of Operations Management* 39:6–22.

Vogel, S., and H. Meyr. 2015. "Decentral Allocation Planning in Multi-Stage Customer Hierarchies". *European Journal of Operational Research* 246(2):462–470.

Yang and Sohn 2021. "Wall Street Journal: Global Chip Shortage 'Is Far From Over' as Wait Times Get Longer". https://www.wsj.com/articles/global-chip-shortage-is-far-from-over-as-wait-times-get-longer-11635413402, accessed 22nd March.

Ziarnetzky, T., and L. Mönch. 2016. "Simulation-Based Optimization for Integrated Production Planning and Capacity Expansion Decisions". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 2992–3003. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ziarnetzky, T., L. Mönch, T. Ponsignon, and H. Ehm. 2019. "Integrated Planning of Production and Engineering Activities in Semiconductor Supply Chains: A Simulation Study". In *Proceedings of the 2019 Winter Simulation Conference*, edited by N. Mustafee, K.-H. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, 2324–2335. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

DR.-ING. VOLKER DÖRRSAM has more than 20 years experience in semiconductor supply chain management. He was heading several operational supply chain teams within the automotive division of Infineon Technologies. Today he is head of the supply chain excellence team to support the operational supply chain teams with digitalization tools and methods. His e-mail address is volker.doerrsam@infineon.com.

JAN-PHILIP ERDMANN currently focuses his research on threshold prediction in ATP-based allocation in semiconductor supply chains. He is studying Business Administration and Engineering with a focus on Mechanical Engineering at RWTH Aachen University. His e-mail address is jan-philip.erdmann@infineon.com.

PATRICK MODER currently conducts research as Ph.D. Candidate focusing on data-informed and automated decision support in semiconductor supply chain when order fulfillment is at risk. He received his Master of Science (Dipl.-Ing.) degree from the Technical University of Munich with a major in Mechanical Engineering and Management. Besides decision support for operations management, he is interested in innovative technologies, such as machine learning or semantic web, and how humans interact with them in a meaningful way. His e-mail address is patrick.moder@infineon.com.