Central European Journal of Operations Research (2024) 32:865–896
https://doi.org/10.1007/s10100-023-00891-z

A rolling horizon planning approach for short-term demand supply matching

Raphael Herding[1,2] · Lars Mönch[1]

Accepted: 8 October 2023 / Published online: 28 October 2023

Abstract
Demand fulfillment and order management are important functions in semiconductor supply chains to interact with customers. In this paper, an iterative short-term demand supply matching (STDSM) algorithm based on mixed-integer linear programming (MILP) is proposed. This approach repromises orders taking into account the finite capacity of the shop floor. Decomposition is used to obtain computationally tractable subproblems. The STDSM approach is applied together with master planning and allocation planning in a rolling horizon setting. A simulation model of a simplified semiconductor supply chain is used for the rolling horizon experiments. The experiments demonstrate that the proposed STDSM scheme outperforms conventional business rule-based heuristics with respect to several delivery performance-related measures and with respect to stability.

Keywords Demand fulfillment · Short-term demand supply matching · Rolling horizon · Semiconductor supply chains · Discrete-event simulation

1 Introduction

The semiconductor industry which manufactures integrated circuits (ICs) is one of the most complex industries in today's world (Mönch et al. 2013). The manufacturing of ICs takes place in a network of frontend (FE) and backend (BE) facilities. A FE facility consists of a wafer fab and a probe/sort area. Starting from a raw wafer, a thin silicon disc, the ICs are produced layer-by-layer on the wafer surface in a wafer fab. The wafers are then sent to a BE facility consisting of an assembly and test (A/T) facility and a final test area.

✉ Lars Mönch
Lars.Moench@fernuni-hagen.de

[1] Department of Mathematics and Computer Science, University of Hagen, 58097 Hagen, Germany

[2] Westphalian University of Applied Sciences, 46397 Bocholt, Germany

Complex process flows in which machines are visited many times by jobs, also called lots in semiconductor manufacturing, are a result of the layer-based manufacturing of ICs. This reentrant behavior results in complex competition for scarce capacity. Long cycle times are common in semiconductor supply chains where the cycle time is the delay between work being released and its emerging as output. Semiconductor supply chains are challenging for existing planning and control approaches and the related information systems (Chien et al. 2011).

Demand fulfillment and order management are important in supply chains (Fleischmann and Meyr 2004; Kilger and Meyr 2015). Commercial advanced planning and scheduling (APS) systems are not appropriate for demand fulfillment in semiconductor supply chains (Chien et al. 2016). This is caused by the large number of products, the complexity of the process flows, the difficulty of capacity modeling due to reentrant flows, the size of the production facilities, and the large-sized supply networks in this industry. It is also shown by Mönch et al. (2018b) that demand fulfillment for semiconductor supply chains is an underresearched area. This is at least partially caused by the fact that demand fulfillment strongly interacts with other planning functions which makes it difficult to study it in a stand-alone manner.

In the present paper, we are interested in proposing a STDSM approach for semiconductor supply chains. Since it is not reasonable to computationally assess the performance of the STDSM approach in isolation, we embed it into a hierarchical approach that contains master planning, allocation planning, release planning, and scheduling. The STDSM approach is based on decomposition that exploits the structure of the semiconductor supply chain. An iterative method is proposed to improve previously made matching decisions. To the best of our knowledge such an approach has not been discussed in the literature yet (cf. Mönch et al. 2018b).

The contribution of this paper is two-fold:

1. We analyze the STDSM planning problem for semiconductor supply chains and propose a corresponding planning approach.
2. The performance of the proposed STDSM approach is assessed in a dynamic and stochastic setting using a rolling horizon scheme based on discrete-event simulation. Incorporating master planning and allocation planning is crucial for this goal since both planning functions provide instructions for the STDSM function.

The paper is organized as follows. In the next section, we describe the problem and discuss related work. The planning approach is presented in Sect. 3. This includes a network-wide allocation planning approach and the STDSM scheme. In Sect. 4, we describe the simulation infrastructure that is used to apply master planning, allocation planning, and the STDSM scheme in a rolling horizon setting. Moreover, the supply chain simulation model and the demand generation scheme are described. The results of simulation experiments are presented and analyzed in Sect. 5. Conclusions and future research directions are discussed in Sect. 6.

2 Problem description and discussion of related work

We start by describing the demand fulfillment function in semiconductor supply chains in Sect. 2.1. We then discuss related work in Sect. 2.2.

2.1 Demand fulfillment in semiconductor supply chains

The following three aspects of the demand fulfillment functionality can be distinguished (Fleischmann and Meyr 2004; Kilger and Meyr 2015; Mönch et al. 2018b):

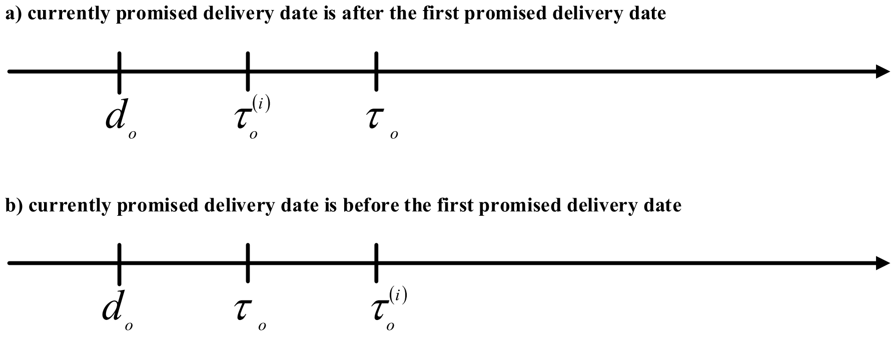
1. Allocation planning
2. Order promising
3. Available to promise (ATP) reallocation and STDSM.

Allocation planning deals with assigning the projected supply of products to customers. We refer to the projected supply as ATP quantities. A distinction is made between committed and uncommitted ATP quantities. ATP quantities can be computed on the capacity planning, the master planning, or the production planning and scheduling level. Order promising is responsible for the ATP consumption by orders. If firm orders arrive they are matched with the corresponding ATP quantities. Three different order promising modes are differentiated:

1. Online order promising: An order is immediately promised after the customer places an order.
2. Batch order promising: All orders placed during the batch interval are simultaneously considered at the end of the batch interval. They are promised at a specific point in time.
3. Hybrid order promising: Online order promising activities are carried out for a certain period of time, followed by a batch promising step afterwards where the previously made promising decisions are confirmed and improved.

The desired delivery date for order o is denoted by d_o. A first promised delivery date $\tau_o^{(i)}$ is chosen for each order o by order promising. The superscript (i) is used to indicate that this is the initially promised delivery date. Moreover, since the promised delivery date of an order can change over time due to periodically performed STDSM activities, we consider the currently promised delivery date τ_o for order o. This delivery date is computed during the last performed STDSM activity. The different types of delivery dates are shown in Fig. 1 where we assume without loss of generality that d_o is before $\tau_o^{(i)}$ and τ_o.

ATP reallocation approaches are responsible for releasing unused committed ATP quotas. All already promised but unfinished orders are considered within a STDSM approach (Fleischmann and Meyr 2004) taking into account the available supply and capacity. STDSM approaches are desirable in semiconductor supply chains due to the long cycle times and the process and demand uncertainty (Mönch et al. 2018b).


a) currently promised delivery date is after the first promised delivery date

b) currently promised delivery date is before the first promised delivery date

Fig. 1 Different delivery dates

A STDSM approach strives to keep the promised delivery dates and to perform manufacturing at the lowest possible cost. Order repromising is required due to high uncertainty and the resulting changes in supply and available capacity. The STDSM function is similar to batch promising, however, all already promised orders compete for the supply and the capacity, while only the orders arriving within the batch interval are considered in batch order promising. The number of orders treated by STDSM approaches is large compared to batch order promising (Geier 2014). Note that in the literature (Fleischmann and Meyr 2004; Geier 2014) the notion of demand supply matching is typically used for the STDSM function described in the present paper. But demand supply matching approaches are also known on a more aggregated, mid-term level in semiconductor supply chains, for instance, the model predictive control approach by Smith and Kempf (2005) and semiconductor-specific master planning approaches (Mönch et al. 2018b). However, orders are not explicitly considered in these approaches. We refer to STDSM when an order-based matching takes place on a short-term level.

The literature for demand fulfillment in semiconductor supply chains is limited (see Sect. 2.2). To the best of our knowledge STDSM approaches in semiconductor supply chains are rule-based taking into account ATP quantities (Herding et al. 2017). This paper contributes to this literature by designing a STDSM approach that considers available capacity in the FE and BE facilities while changing the current promised delivery dates of already promised orders as little as possible. Because of the large size of semiconductor supply chains, the proposed STDSM approach is based on decomposition.

The research questions addressed in this paper can be summarized as follows:

1. What are the design principles of a STDSM approach that is able to take into account process conditions of semiconductor supply chains?
2. How can the STDSM approach be embedded into a hierarchical approach for planning and control of semiconductor supply chains?
3. What is the performance of the STDSM approach with respect to solution quality and computing time relative to a conventional rule-based repromising approach that is only based on ATP quantities but not on available supply and capacities?

2.2 Related work

2.2.1 Demand fulfillment in semiconductor supply chain planning systems

Several early papers mention demand fulfillment-related subsystems of semiconductor supply chain planning systems. For instance, a module of the IMPReSS production planning system at Harris Corporation calculates product availability for the quotation and order entry system (Leachman et al. 1996). Requirement and system specification efforts are described by Soares et al. (2000) for an order promising module of a decision support system for semiconductor supply chains, but computational results are not reported. The PROFIT planning system implemented at IBM Semiconductor contains an ATP module (Lyon et al. 2001). Some semiconductor companies use commercial APS systems for demand fulfillment tasks in their daily business (Chien et al. 2016). A capable-to-match (CTM) algorithm for the APS system SAP APO is discussed by Kallrath and Maindl (2006). The CTM approach is similar to the STDSM functionality. However, details are not provided for all these systems that provide demand fulfillment functionality.

2.2.2 Allocation planning

Semiconductor-specific allocation planning approaches are rare (cf. Mönch et al. 2018b). However, there are a few papers for other industrial domains that can be extended towards semiconductor supply chains. An allocation planning approach for the lighting industry is proposed by Meyr (2009). The approach first segments customers with respect to their importance and profitability into different priority classes. ATP quantities are allocated to these classes based on short-term demand information. The objective is profit maximization. Several ATP consumption strategies are tested. Seitz et al. (2020) extend the allocation planning approach of Meyr (2009) by exploiting the known demand forecast bias of customers. Using data from a large semiconductor manufacturer, it is shown by designed experiments that average stock levels are reduced and the overall service level is increased. This is especially true for customers that provide truthful forecasts. An allocation planning model similar to the model of Meyr (2009) is proposed by Babarogić et al. (2012). Customers are assigned to priority groups based on the size of their orders. The objective consists in maximizing the service level. Computational examples from the fast-moving consumer goods industry are used. An allocation planning procedure for an assemble-to-order (ATO) supply chain is proposed by Chen and Dong (2014). Multiple facilities producing components that are used in various final products are assumed. Assembly operations are used to produce the end products. The proposed allocation planning approach considers the finite capacity of the different facilities. A demand fulfillment system for semiconductor foundries in Taiwan is described by Chiang and Hsu (2014). An allocation planning component is proposed that respects highly aggregated bottleneck capacities from capacity planning. Moreover, LP-based order promising models are designed. A period-based allocation review mechanism is proposed that reallocates unused ATP quantities. The allocation planning model of Chiang and Hsu (2014) is investigated by Framinan and Perez-Gonzalez (2016) with inaccurate and biased forecast and the situation that only a certain fraction of the overall capacity may be allocated to specific products and customers. An online order promising approach is taken where the arrival of firm orders is simulated. The simulation experiments show that the allocation planning scheme is sensitive to inaccurate and biased forecasts. Caps on the capacity to be allocated can be seen as a strategy to deal with forecast inaccuracy. An allocation planning approach for semiconductor manufacturing is proposed by Mousavi et al. (2019). The service level and the reserved buffer stock are considered in a bi-criteria setting. A MILP is used to make allocation decisions. However, different objectives for allocation planning are considered in the present paper in a multi-facility setting which is different from Mousavi et al. (2019). The single-facility, single-product allocation planning approach of Meyr (2009) is extended towards multiple products and alternative facilities by Azevedo et al. (2016). In the present paper, we will use a multi-facility procedure similar to the allocation approach by Meyr (2009) and Azevedo et al. (2016).

2.2.3 Demand fulfillment in the thin-film-transistor liquid–crystal display industry

Another stream of related work deals with demand fulfillment in thin-film-transistor liquid–crystal display (TFT-LCD) manufacturing which is close to wafer fabrication but much simpler. An ATP model for computing a promised delivery date for each order is proposed by Jeong et al. (2002). A capable to promise (CTP) model is designed that determines the unused capacity of the shop floor for a module assembly schedule. However, this problem is different from our problem since we use the available capacity for repromising orders. Tsai and Wang (2009) propose a three-phase approach for a TFT-LCD ATO manufacturing setting. Orders are assigned to module plants in a first phase taking into account aggregated capacity and material availability. In a second phase, the ATP allocation to orders in single module plant is considered for a given order due date. Orders that cannot be allocated in the two phases are reallocated to all module plants. This approach is similar to the proposed STDSM approach, but instead of using a fixed desired delivery date d_o we propose an iterative approach for repromising orders based on time windows of increasing length. Experiments in a rolling horizon setting are not described by Tsai and Wang (2009). Therefore, important measures related to the first promised delivery date cannot be computed. Lin et al. (2010) design a batch order promising approach. Alternative bill of materials and multiple quality grades are taken into account. The impact of the batch interval length on profit is studied. While the approach addresses important features of semiconductor supply chains, it is a batch order promising approach which is different from the STDSM function.

2.2.4 Rolling horizon approaches for demand fulfillment

A scalable infrastructure for supply chains is applied to batch order promising by Zhao et al. (2003) on a conceptual level. The need for rolling horizon approaches for assessing demand fulfillment is conceptually discussed by Chen et al. (2008). A STDSM approach is proposed by Geier (2014) for a computer manufacturer. It is integrated with order promising in a rolling horizon setting, while feedback from the shop floor is considered. The STDSM approach proposed in the present paper is different since we compute the supply for BE facilities based on FE production planning. Moreover, we use an iterative approach that extends the delivery time windows of the orders. Seitz and Grunow (2017) propose an order promising approach that exploits product and process flexibility typical for semiconductor supply chains. ATP information is determined by rolling horizon production planning, but feedback from the shop floor is not taken into account. The interaction of order promising and master production scheduling for a ceramic tile company is studied by Alemany et al. (2018). However, the integrated approach is not assessed using a rolling horizon scheme and simulation. An infrastructure for simulation-based performance assessment of demand fulfillment is proposed by Herding et al. (2017). However, only some preliminary computational results for the interaction of master planning and rule-based online order promising and repromising are presented in this paper.

To the best of our knowledge, there is no approach described in the literature that covers the interaction of master planning, allocation planning, and order promising and repromising for semiconductor supply chains. Optimization-based STDSM approaches are not considered so far in the literature for semiconductor supply chains. Assessing this interaction in a rolling horizon setting under process and demand uncertainty is highly desirable.

3 Planning approach

We discuss the overall planning approach and the underlying assumptions in Sect. 3.1. Master planning and allocation planning as prerequisite for the STDSM are briefly sketched in Sect. 3.2. The proposed STDSM models are presented in Sect. 3.3. The reference approach and the remaining planning and control functions are discussed in Sect. 3.4.

3.1 Assumptions and overall approach

A semiconductor supply chain consists of several FE and BE facilities. The probed wafers are stored in die banks (DBs) that serve as decoupling points between FE and BE. Distribution centers (DCs) are responsible for decoupling BE facilities and customers. Each FE and BE facility consists of machine groups which contain machines that provide the same functionality. We refer to machine groups as work centers in the rest of this paper. We start by describing different product aggregates, i.e. a grouping of products based on certain criteria, to characterize the supply. The internal view of sellable products is given by finished products (FPs) that are available at the DCs. FPs contain information regarding which FE and BE facilities produce the product. When the FE facility of a product is known, but the BE facility and the DC are not yet determined, the product is represented by a DB representative (DREP) in the supply picture provided by master planning. DREP products are available at the different DBs. Finally, the fabrication position (FPOS) aggregate is used to represent the FE level in the supply picture offered by master planning. The DB of FPOS aggregates is not determined yet. In this paper, we differentiate between orders that are fulfilled by FP, DREP, and FPOS product aggregates. The structure of the considered semiconductor supply chains including the different product aggregates is shown in Fig. 2.

The proposed STDSM approach is based on the following assumptions:

1. Process and product flexibility exist in semiconductor supply chains (Lyon et al. 2001; Mönch et al. 2018a). The former means that a single production process can be used to manufacture several products, and the latter refers to the possibility to produce several products from one predecessor product. Therefore, the general product master data is given by a graph. However, for ease of exposition, we assume a 1:1 relationship between FP, DREP, and FPOS in the present paper.
2. Only the capacity of the bottleneck work centers is taken into account by master planning to reduce the size of the resulting MILP model (see Sect. 3.2).
3. The master planning formulation and the different planning models of the STDSM approach are based on exogenous lead times that are an integer multiple of the period length. Lead times are estimates of the cycle time (CT).
4. Supply is given by master planning which determines what quantities of the considered semi-finished and finished products have to be completed in which FE and BE facility of the considered supply chain in which period of the planning horizon (Mönch et al. 2018b). More specifically, the supply for FPOS products is given by the quantity of a product to be completed at the end of a given period in a given FE facility. Moreover, we consider supply for DREP products which is also computed by master planning.
5. Splitting of orders for partial order promising is not considered in this paper.

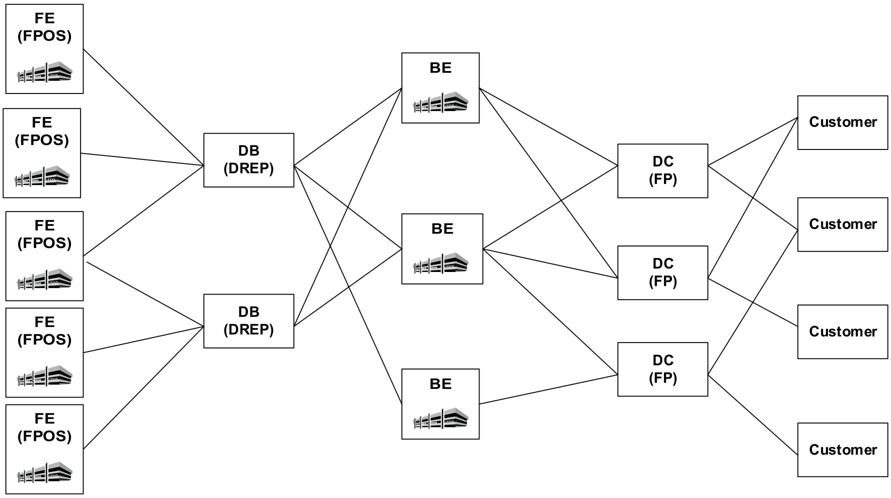


Fig. 2 Main entities in semiconductor supply chains

Next, we describe both the allocation and the STDSM planning approach. We differentiate subproblems that refer to the BE and the FE yielding to decision models called BE STDSM and FE STDSM (these LP or MILP models will be described in the Sect. 3.3). The proposed planning approach consists of the following steps:

1. We use the FE STDSM-NO model with demand that is obtained from the supply determined by the master planning function. The supply is given by the quantity of a product to be completed at the end of a period in a given FE facility. No orders (NO) have to be repromised in this step. Since the FE STDSM-NO model considers capacity constraints for all work centers, not just for the bottlenecks (see Sect. 3.3), it is able to provide more accurate supply than the master planning model. Here, the supply is given by the output quantities of the FE STDSM-NO model for each product, period, and FE facility. The supply provided in this step is required for allocation planning (Step 2) and for the BE STDSM model (Step 3).
2. The supply computed in Step 1 is used together with demand information to derive allocated ATP (AATP) quantities, i.e., scarce ATP quantities are assigned to customers.
3. Based on the supply computed in Step 1 and the AATP quantities from Step 2, the BE STDSM model aims at repromising all orders at the first promised delivery date $\tau_o^{(i)}$ on the DC level. If the repromised date of an order o is different from $\tau_o^{(i)}$, this order is reconsidered in Step 4. Orders that are repromised at $\tau_o^{(i)}$ are not considered anymore.
4. The goal of this step is to repromise orders o with a repromised date different from $\tau_o^{(i)}$ in Step 3 as close as possible to $\tau_o^{(i)}$ using the FE STDSM-O model where the acronym O indicates that orders have to be repromised. Instead of using the original $\tau_o^{(i)}$ values, we modify them by subtracting the BE lead time measured in periods. Estimates for the BE lead time are obtained by simulation. Orders from Step 3 and AATP quantities from Step 2 serve as input for this step. Orders that are not repromised in Step 3 at $\tau_o^{(i)}$ can be repromised in Step 4 even if this decision might lead to a changing τ_o value. This is motivated by the observation that repromising orders is crucial, whereas the concrete τ_o value is less important. Step 4 can be iteratively repeated by increasing the allowed delivery time window for each order (see below for a description).

Since it is more likely that orders are repromised when they are equipped with a delivery time window covering multiple periods rather than a single allowed period, namely $\tau_o^{(i)}$, we choose the start and end date of a delivery time window $\left[e_o, f_o\right]$ for order o as follows:

$$e_o^r :=\max\left(\tau_{\mathrm{o}}^{(\mathrm{i})}-k(r),\ 1\right) \tag{1}$$

$$f_o^r :=\min\left(\tau_{\mathrm{o}}^{(\mathrm{i})}+l(r), T\right), \tag{2}$$

where T is the length of the planning window measured in number of periods, r the iteration counter, and k and l are non-decreasing functions with argument r that have to be specified in a concrete situation. Note that the maximization with 1 in (1) and the minimization with T in (2) ensure that $e_o^r \geq 1$ and $f_o^r \leq T$, i.e. that the planning window is respected by the start and end date of the delivery time window. This results in more restricted time windows for orders with a $\tau_o^{(i)}$ value close to the beginning or end of the planning window. Due to the rolling horizon approach this is not crucial for orders with a $\tau_o^{(i)}$ value close to the end of the planning window. Moreover, it will be penalized in the objective function of the FE STDSM-O model when orders with a $\tau_o^{(i)}$ value close to the beginning of the planning window cannot be repromised within the given time window. We start by iterations where we use increasing $k(r)$ values, followed by iterations where we increase $l(r)$ since orders that are fulfilled before $\tau_o^{(i)}$ are considered as inventory that can be delivered at $\tau_o^{(i)}$. If orders are repromised within a single iteration, they are not considered anymore in the following iterations. The corresponding repromising decisions are incorporated into the FE STDSM-O model by fixing the values of the related decision variables. This allows respecting previously made order repromising decisions. The overall STDSM approach including allocation planning activities is summarized in Fig. 3.

The FE and BE STDSM MILP instances can be solved individually for each single FE and BE facility since supply is provided by master planning for each single facility. This is indicated by individual boxes for the different FE facilities (indicated by FE1, …, FEm) in Step 1 and Step 4. However, we need a mechanism to assign orders to the different facilities. In the present paper, orders are randomly assigned to the FE facilities where all facilities have the same probability to be selected.

The BE facilities are much smaller with respect to the number of work centers and number of process steps in the routes (Mönch et al. 2013). Therefore, solving

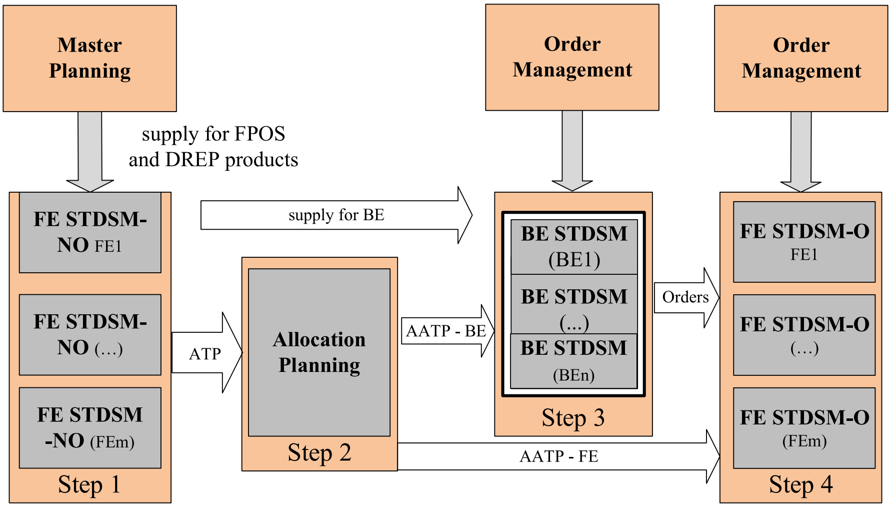


Fig. 3 STDSM planning approach

a simultaneous BE STDSM MILP instance for all BE facilities is possible. This is indicated by the surrounding frame for the BE facilities (Step 3) in Fig. 3.

Note that the proposed planning approach is somehow similar to the FE- and BE-based production planning decomposition procedure used in the decision support system IMPReSS (Leachman et al. 1996). However, orders are not considered in IMPReSS in contrast to the present paper. Next, we will describe the different ingredients of the proposed planning approach.

3.2 Master planning and allocation planning

A generic planning window of finite length T that consists of equidistant periods $t = 1, \ldots, T$ is assumed for all planning formulations in the rest of the paper. The master planning formulation extends the model by Ponsignon and Mönch (2012) for several FE facilities to the situation that BE facilities are included. The LP model assumes fixed integer lead times for both FE and BE facilities. For simplicity reasons, all products have the same lead time. Capacity constraints are only taken into account for FE/BE bottleneck work centers to reduce the model size. The model is formulated for a set of demand classes I with different priorities (Leachman et al. 1996). For instance, previously confirmed customer orders form the highest priority class, replenishment to target inventory levels is the second important class, sales forecasts discounted by historical forecast errors is the third most important class, and the rest of sales forecast, i.e. the risky portion, forms the least important demand class. The main decision variables y_{gjt}^{FE},y_{gjt}^{BE}, I_{gt}^{DB}, I_{gt}^{DC} are for the number of lots of product g to be completed at the end of period t in FE/BE facility j and the DB/DC inventory level for product g in period t, respectively. The full model can be found in Herding and Mönch (2021) for the sake of completeness.

Allocation planning is responsible for allocating scarce ATP quantities, obtained by Step 1 of the proposed planning approach, to different customers. A multi-facility version of the allocation planning approach by Meyr (2009) is presented in Appendix A of the electronic supplement. The objective function of the resulting LP is the difference of the sum of the weighted AATP quantities and a penalty term for not meeting the given minimum ATP quantities for the different customers. The main FE decision variables are the $aatp_{jcigt\tau}^{FE}$ variables which represent ATP for demand class i of product g in FE facility j, available at the begin of period t, allocated to demand for customer c which is due in period τ. The main BE decision variables are the $aatp_{cigt\tau}^{BE}$ variables which model the ATP for demand class i of product g for all BE facilities, available at the begin of period t, allocated to demand for customer c which is due in period τ. In order to use the AATP quantities in the order repromising approach, we have to differentiate between the AATP quantities for the FE and BE repromising approaches. We set $aatp_{cgt\tau}^{FE} := aatp_{c1gt\tau}^{FE}$ and $aatp_{cgt\tau}^{BE} := aatp_{c1gt\tau}^{BE}$ for the FE and BE approach, respectively, where demand class 1 refers to confirmed orders. The AATP quantities of demand class 2 that refers to forecasted demand are used in the online order promising procedure that will be described in Sect. 3.4.

3.3 FE STDSM and BE STDSM models

We start by formulating the FE STDSM model variants, i.e. the FE STDSM-O and the FE STDSM-NO. They are based on the following sets and indices, decision variables, and parameters.
Sets and indices.

t, τ:	Period index
$g \in G$:	Product index for set of all products
$j \in F$:	Facility index for set of all FE facilities
$k \in K^{FE}(j)$:	Work center index for set of all work centers of FE facility j
l:	Operation index
l^*:	Last operation of the route for product g in facility j
$n \in N$:	Product type index for set of all product types for FE facilities, $N = \{DREP, FPOS\}$
o:	Order index
$c \in \Omega$:	Customer index for set of all customers
$O^{FE}(g,j)$:	Set of all operations of product g in facility j
$O^{FE}(g,j,k)$:	Set of all operations of product g on machines of work center k of facility j
A_{gc}:	Set of all orders of product g for customer c
A_g:	Set of all orders of product g

Decision variables.

Y^{FE}_{jgtl}:	Quantity of product g in facility j completing operation l in period t
Y^{FE}_{jgt}:	Output of product g in facility j in period t from the last operation of its routing
X^{FE}_{jgt}:	Quantity of product g released into the first work center of facility j in its routing in period t
W^{FE}_{jgt}:	WIP of product g in facility j at the end of period t
$S^{n}_{ot\tau}$:	1 if order o is completed by product type n in period $t \leq \tau$, 0 otherwise
I^{DB}_{gt}:	DB inventory of product g at the end of period t
B_{gt}:	Backlog of product g at the end of period t

Parameters.

q_o:	Size of order o (in wafers)
π^n_o:	Unit revenue of order o assigned to product type n
$l_{o\tau}$:	Unit penalty value if order o is not repromised for period τ
e_o:	Earliest delivery date of order o
f_o:	Latest delivery date of order o
h_{gt}:	Unit DB holding cost for product g in period t
ω^{FE}_{jgt}:	Unit WIP cost of FE facility j for product g in period t
b_{gt}:	Unit backlog cost for product g in period t
$Y^{FE(i)}_{jgt}$:	Initial quantity (in wafers) of product g in facility j to be completed at the end of period t
C_{jkt}:	Available capacity of work center k of facility j during period t
α_{jgl}:	Processing time of operation l of product g in facility j

L_{gl}: FE lead time (in number of periods) for product g from release of the raw material to the completion of operation l

$aatp^{FE}_{cgt\tau}$: ATP quantity allocated to confirmed orders of product g for customer c due in period τ, available at the begin of period t

$\tilde{S}_{gt}$: Requested quantity of product g to be completed in period t (supply for FPOS products from master planning)

Next, the FE STDSM-O model is formulated as follows:

$$\max \sum_{g\in G}\sum_{t=1}^{T}\left[\sum_{o\in A_g}\sum_{n\in N}\sum_{\tau=e_o}^{f_o} q_o\left(\pi_o^n - l_{o\tau} + l_{o,T+1}\right) S^n_{ot\tau} - \sum_{j\in F}\omega^{FE}_{jgt}W^{FE}_{jgt} - h_{gt}I^{DB}_{gt}\right] \tag{3}$$

subject to

$$W^{FE}_{j,g,t-1} + X^{FE}_{jgt} - Y^{FE}_{jgt} = W^{FE}_{jgt} \quad j\in F, t=1,\dots,T, g\in G \tag{4}$$

$$\sum_{j\in F} Y^{FE}_{jgt} + \sum_{j\in F} Y^{FE(i)}_{jgt} + I^{DB}_{g,t-1} - I^{DB}_{gt} = \sum_{\tau=1}^{T}\sum_{o\in A_g}\sum_{n\in N} q_o S^n_{ot\tau} \quad t=1,\dots,T, \quad g\in G \tag{5}$$

$$\sum_{g\in G}\sum_{l\in O^{FE}(g,j,k)} \alpha_{jgl}Y^{FE}_{jgtl} \le C_{jkt} \quad j\in F, t=1,\dots,T, k\in K^{FE}(j) \tag{6}$$

$$Y^{FE}_{jgtl} = X^{FE}_{j,g,t-\lfloor L_{gl}\rfloor} \quad j\in F, t=1,\dots,T, g\in G, l\in O^{FE}(g,j) \tag{7}$$

$$Y^{FE}_{jgt} = Y^{FE}_{jgtl*} \quad j\in F, t=1,\dots,T, g\in G \tag{8}$$

$$\sum_{\tau=1}^{T}\sum_{o\in A_g} q_o S^n_{ot\tau} \le \sum_{j\in F} Y^{FE}_{jgt} + \sum_{j\in F} Y^{FE(i)}_{jgt} \quad n = FPOS, t=1,\dots,T, \quad g\in G \tag{9}$$

$$\sum_{\tau=1}^{T}\sum_{o\in A_g} q_o S^n_{ot\tau} \le I^{DB}_{g,t-1} \quad n = DREP, t=1,\dots,T, g\in G \tag{10}$$

$$\sum_{n\in N}\sum_{t=1}^{T}\sum_{\tau=e_o}^{f_o} S^n_{ot\tau} \le 1 \quad g\in G, o\in A_g \tag{11}$$

$$\sum_{n\in N}\sum_{t=1}^{T}\sum_{\tau\notin[e_o,f_o]} S^n_{ot\tau} = 0 \quad g\in G, o\in A_g \tag{12}$$

$$\sum_{n\in N}\sum_{o\in A_{gc}} q_o S^n_{ot\tau} \le aatp^{FE}_{cgt\tau} \quad g\in G, c\in \Omega,\ t,\tau\in\{1,\dots,T\} \tag{13}$$

$$X^{FE}_{jgt}, Y^{FE}_{jgtl}, Y^{FE}_{jgt}, W^{FE}_{jgt}, I^{DB}_{gt}, B_{gt} \ge 0 \quad j\in F, t=1,\dots,T, g\in G, l\in O^{FE}(g,j) \tag{14}$$

$$S^n_{ot\tau}\in\{0,1\} \quad t,\tau\in\{1,\dots,T\}, g\in G, o\in A_g, n\in N. \tag{15}$$

The objective (3) seeks to maximize the profit, i.e. the difference of the revenue of the repromised orders and the sum of costs. The first term of (3) represents the difference of the revenue and the cost for repromising certain orders in a period different from $\tau^{(i)}_o$. We use

$$l_{o\tau} := \begin{cases} \alpha_o\left(\tau^{(i)}_o - \tau\right), & \text{if } e_o \le \tau \le \tau^{(i)}_o \\ \beta_o\left(\tau - \tau^{(i)}_o\right), & \text{if } \tau^{(i)}_o < \tau \le f^{(i)}_o \\ \beta_o\left(T+1-\tau^{(i)}_o\right), & \text{otherwise} \end{cases} \tag{16}$$

for given order-specific quantities $1 \le \beta_o \le \alpha_o$. This setting ensures that the model prefers repromising orders within the time window at or before $\tau^{(i)}_o$, followed by repromising them after $\tau^{(i)}_o$ within the time window. Orders that cannot repromised within the time window are artificially repromised at period $T+1$. We penalize this in the objective function by adding a term including $l_{o,T+1}$ to the objective function. The second term models WIP costs. The cost for holding inventory at the DB is given by the third term.

Constraints (4) represent the WIP balance for each FE facility. Constraints (5) are inventory balance equations. The capacity restrictions for each work center are ensured by constraints (6). The C_{jkt} values are adjusted in such a way that the initial WIP, represented by $Y^{FE(i)}_{jgt}$ is taken into account. Integer lead times that are a multiple of the period length are incorporated into the model by the input–output relation constraints (7). Simulation is used to determine appropriate waiting time estimates for computing operation-specific lead times L_{gl} in a recursive manner. The lead time is then obtained by rounding down the non-integer estimates obtained from the recursion (cf. Kacar et al. 2016; Missbauer and Uzsoy 2020). Constraint set (8) sets the values of the decision variables Y^{FE}_{jgt} to Y^{FE}_{jgtl*}. Constraint sets (9) and (10) model the balance for order repromising. The orders can be repromised by product types DREP and FPOS. Here, the amount of DREP products in a given period is determined by the values of the $I^{DB}_{g,t-1}$ decision variables whereas the amount of FPOS products is represented by the values of the Y^{FE}_{jgt} decision variables. This means that constraints (9) ensure that the amount of repromised orders is not larger than the amount of completed lots that belong to FPOS, whereas constraints (10) model the same for DREP. The constraints (11) and (12) make sure that an order can only be repromised within its time window, whereas constraint set (13) ensures that the amount of orders per product and customer is not larger than the ATP quantities that are allocated to customer-specific demand that is due in period τ. The range of the decision variables is modeled by the constraints (14)–(15).

Different product types N are used in the FE STDSM-O model to support repromising on the DREP and the FPOS level, respectively. The revenue of orders is selected as $\pi_o^{DREP} > \pi_o^{FPOS}$ to make sure that if possible, orders are repromised first as DREP before FPOS is used.

When we have $A_g = \emptyset$ for all $g \in G$, i.e., orders are not considered in the formulation, we call the resulting model FE STDSM-NO (see Sect. 3.1). It is obtained from the FE STDSM-O model based on the following changes. First, the objective function (3) is replaced by

$$\sum_{g\in G}\sum_{t=1}^{T}\left[\sum_{j\in F}\omega_{jgt}^{FE}W_{jgt}^{FE} + h_{gt}I_{gt}^{DB} + b_{gt}B_{gt}\right] \tag{17}$$

which has to be minimized. In (17), the third term models backlog cost. Moreover, the constraint sets (5), (9), and (10) are replaced by

$$\sum_{j\in F} Y_{jgt}^{FE} + \sum_{j\in F} Y_{jgt}^{FE(i)} + I_{g,t-1}^{DB} - B_{g,t-1} + B_{gt} - I_{gt}^{DB} = \tilde{S}_{gt} \quad t = 1,\dots,T, \quad g \in G. \tag{18}$$

Constraints (18) are inventory balance equations. Constraint sets (11)–(13) are obsolete. Note that we take the I_{g0}^{DB} quantity, i.e. the initial DREP supply, from master planning. The right-hand side $\tilde{S}_{gt}$ is the supply for FPOS products, i.e. $\tilde{S}_{gt} := \sum_{j\in FE} y_{gjt}^{FE}$. Note that backlog is allowed in the FE STDSM-NO model to avoid infeasibilities due to limited capacity.

The BE STDSM model is similar to the FE STDSM-O model (3)–(15). Therefore, we only introduce the main decision variables and the constraints that are different from the FE-STDSM-O model. X_{jgt}^{BE} is the quantity of product g released into the first work center of BE facility j in its routing in period t whereas Y_{jgt}^{BE} is the output of product g in facility j in period t from the last operation of its routing. The binary decision variable $S_{ot\tau}$ is 1 if order o is completed in period $t \leq \tau$ and 0 otherwise. The objective function of the BE STDSM model is similar to (3). An FE inventory balance equation similar to (18) models the supply obtained from Step 1. Moreover, we have a BE inventory balance equation where the right-hand side is given by the sum of the sizes of the orders that are repromised in a given period. In addition, there is a constraint set similar to (13). Here, we have to replace the right-hand side of (13), i.e. $aatp_{cgt\tau}^{FE}$, by $aatp_{cgt\tau}^{BE}$. The complete model can be found in Appendix B of the electronic supplement.

Due to the binary decision variables, large-sized instances of both MILP models are hard to solve. This can be seen from Appendix C of the electronic supplement where we show that both the FE STDSM-O and the BE STDSM problems are NP-hard. Therefore, the FE STDSM-O and the BE STDSM models are solved by a heuristic time-based decomposition technique proposed by Brahimi et al. (2015) for a single-stage order acceptance model. This method is based on a decision interval with a length of α periods and a frozen interval that consists of β periods. Order repromising decisions are modeled by binary decision variables in the decision interval, whereas the binary decision variables are relaxed in the sense that they can take values from $[0, 1]$ in the rest of the planning window. Moreover, the values of the already selected binary decision variables are fixed in the frozen interval.

3.4 Remaining planning and control functions and reference approach

Next, we briefly describe the remaining planning and control functions that are used in all conducted computational experiments. We start by sketching the online order promising (OOP) algorithm. A backward search is performed to find AATP quantities of the forecasted demand class to fulfill order o in periods at or before d_o. A forward search in periods after d_o is carried out if not enough AATP is found during the backward search. Both the backward and the forward search initially strive to find ATP at the DC level. If ATP for an order cannot be fulfilled at the DC level, the algorithm looks for ATP at the DB level, BE lead time periods before. If ATP is still missing, the OOP scheme looks for ATP at the WIP level. The OOP algorithm is similar to the ATP search procedure described by Kilger and Meyr (2015), for details we refer to Herding et al. (2017).

Lot releases are determined by backward termination, a simple production planning approach, for each FE and BE facility based on product-specific quantities that are computed by master planning using lead time information. Waiting time estimates that are a multiple of the processing time are incorporated into the backward termination scheme. Scheduling is carried out using the distributed shifting bottleneck heuristic (DSBH) proposed by Mönch and Drießel (2005). Moreover, lots on the execution level have to be assigned to orders to fulfill them. Following the lot-to-order matching procedure by Knutson et al. (1998), all orders are randomly assigned to a specific FE and BE facility in a first stage. All facilities have the same probability to be chosen. The sequence in which the orders are considered in both stages is determined based on the q_o and τ_o values of order o, i.e., the index $w_o q_o / \tau_o$ is used. Here, w_o is a weight associated with order o. The w_o values are used to express the importance of order o. Orders with tight τ_o values, large q_o quantities, and large weights w_o are preferred. The existing lots in a wafer fab are used to fill the orders following the first fit decrease heuristic from bin packing (Dowsland and Dowsland 1992). The lots are considered in non-decreasing order of their due dates.

We continue by briefly describing the main ideas of the reference approach for the STDSM function, the rescheduling batch run (RBR). It is based on several repromising rules that control the search for AATP quantities similar to the logic used in the OOP scheme (see above). The repromising rule ALL_ON_TIME only repromises an order if the entire order quantity is available at the currently promised delivery date τ_o. Repromising rule ALL only repromises an order if the entire order quantity is available even if this results in a postponement of an order delivery. It collects order quantities until the entire quantity is available. The repromising date is the date when the entire order quantity is available. The RBR procedure can be summarized as follows:

RBR procedure

1. Determine the set of orders O to be repromised. Initialize the set of all already considered orders by $R := \emptyset$. Sort the order set O in a non-decreasing order with respect to the τ_o values.
2. Let o^* be the first order in the sorted list derived from the set O.
3. Apply the ALL_ON_TIME rule by checking whether the entire order quantity of order o^* is available at the DC at τ_{o^*} or not. If yes, go to Step 8.
4. Search for additional order quantities at the DC in the periods at or before τ_{o^*} until the entire order quantity of order o^* is obtained or the beginning of the planning window is reached.
5. If still quantities to be repromised are left for order o^*, search for additional order quantities at the DB at or before τ_{o^*} until the entire order quantity of order o^* is obtained or the beginning of the planning window is reached.
6. If again quantities to be repromised are left, search for additional order quantities in future periods at the DC until the entire order quantity of order o^* is obtained or the end of the planning window is reached.
7. If still quantities to be repromised are left, search for available order quantities in future periods at the DB until the entire order quantity of order o^* is obtained or the end of the planning window is reached.
8. Update $O := O \backslash \{o^*\}$ and $R := R \cup \{o^*\}$. If $O \neq \emptyset$ go to Step 2.
9. Try to improve the repromising decisions by performing a cross-confirmation run (CCR) (described next).

Note that the Steps 4–7 represent the ALL rule. The CCR procedure aims to improve the repromised dates with respect to the desired delivery dates d_o. Therefore, the procedure starts by freezing the remaining supply. All repromised orders are made available again. Based on this newly available supply, the CCR tries to improve the repromised delivery dates by performing again the RBR procedure with a modified Step 1 where the orders are sorted with respect to d_o instead of τ_o. If an improvement is possible, the improved order confirmation date is set, otherwise the previously repromised date is used. We illustrate the RBR by an example shown in Fig. 4.

A planning window of length $T = 5$ periods is considered. The order quantity is $q_{o^*} = 1000$ units. At τ_{o^*}, we have only 200 units available. Hence, the ALL_ON_TIME rule is not applicable (Step 3). In addition to the 200 units at τ_{o^*} we find another 100 units in the period before τ_{o^*} (Step 4). We search on the DB before or at τ_{o^*} and find 100 units in the period before τ_{o^*} (Step 5). Here, for the sake of simplicity we assume that the production in the BE facilities can be carried out in less than a single period. Another 100 units are found in the periods after τ_{o^*} at the DC by carrying out Step 6. In Step 7, we find the remaining 500 units at the DB. Overall, the order o^* is repromised for period 5 while we have $\tau_{o^*} = 3$.

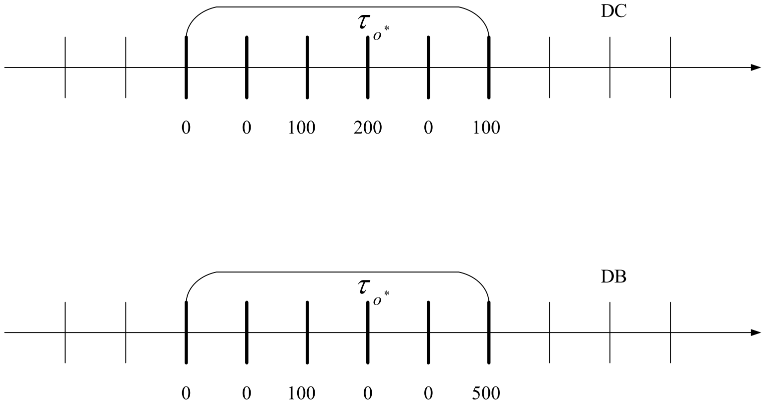


Fig. 4 Example for the RBR procedure (without the CCR)

Note that the STDSM approach is based on decomposition according to the physical structure of the underlying supply chain, i.e., optimization models are solved for the different nodes of the supply chain or groups of them. It is not obvious which one of the two decomposition approaches, i.e. the STDSM or the RBR procedure, is better. It is well-known that planning problems for large-scaled semiconductor supply chains can only be tackled by decomposition (Fordyce et al. 2011). Either the STDSM or the RBR scheme is used in the simulation experiments whereas the remaining planning functions, i.e., master planning, allocation planning, release planning, scheduling, lot-to-to-order matching, and the OOP procedure are the same for both situations. The different planning and control modules are summarized in Table 1.

4 Rolling horizon approach

The simulation infrastructure and the applied simulation model is discussed in Sect. 4.1. The demand and order generation scheme is described in Sect. 4.2.

4.1 Simulation infrastructure and supply chain simulation model

The performance of the proposed STDSM approach can only be reasonably assessed in a rolling horizon manner since several planning functions are applied in different frequency and the value of STDSM decisions must be evaluated based on global, i.e. supply chain-wide performance measures. Discrete-event simulation is crucial for implementing rolling horizon schemes in a risk-free environment due to the fact that the dynamics and the uncertainty of the supply chain can be covered.

The simulation infrastructure proposed by Herding et al. (2017) is extended for the rolling horizon experiments carried out in the course of the research for this paper. It contains a planning, control, and execution level. According to Missbauer and Uzsoy (2022), the purpose of the planning level is to coordinate the material

Table 1 Characteristics of the Different Planning and Control Modules

Planning/control module	Objective function	Input from	Output to
Master planning	Profit	Demand planning	STDSM, lot release planning
STDSM			
FE STDSM-NO	Cost	Supply from master planning for FPOS and DREP products	Supply to allocation planning and to BE STDSM
BE STDSM	Profit based on orders	Supply from FE STDSM-NO	Orders to FE STDSM-O
FE STDSM-O	Profit based on orders	Orders from BE STDSM, supply from FE STDSM-NO	Repromised orders
RBR	–	Already repromised orders and orders from OOP	Repromised orders
Allocation planning	Sum of weighted ATP quantities	Supply from FE STDSM-NO	AATP to FE STDSM-O, OOP
Lot release planning	–	Master planning	Scheduling
OOP	–	AATP from allocation planning	Promised orders for STDSM
Lot-to-order matching	–	Orders from STDSM	Orders for DCs
Scheduling	Total weighted tardiness of the lots	Lots from lot release planning	Dispatching

flow across time and space, while (production) control deals with the monitoring of the progress of work through the supply chain or production facility to identify and address deviations from the plan as they occur. This means that production control is related to lots that are already released on the shop floor. The order penetration point is at the interface between planning and control. A discrete-event simulation model represents the execution level of the semiconductor supply chain, while the master planning, production planning, allocation planning, STDSM, lot-to-order matching, and OOP procedures are located on the planning level. The corresponding planning models are populated with data using feedback from the execution level. The control level is responsible for computing release and dispatching instructions and performing order management activities. The backbone of the infrastructure is a blackboard-type data layer in the memory of the simulation computer. The data layer is coded in the C++ programming language. Business objects such as orders, lots, machines, and routes are stored in the data layer. These objects are updated in an event-driven manner using notification functions of the commercial simulation software AutoSched AP 9.3. A stop and go approach is applied for the rolling horizon simulation experiments under which the simulation engine stops at the beginning of a planning epoch to execute the different planning functions. The simulation continues after a plan is computed by the corresponding planning approach. The overall principle of the rolling horizon approach is shown in Fig. 5.

The first two planning epochs of the master planning and production planning function are indicated in the figure by vertical lines that are blue colored. The same is also shown for allocation planning, the STDSM/RBR procedure, and lot-to-order matching (green colored). We observe that the frequency of applying the different planning functions is different. For instance, allocation planning is more frequently used than master planning, whereas the OOP procedure is applied whenever an order is placed by a customer.

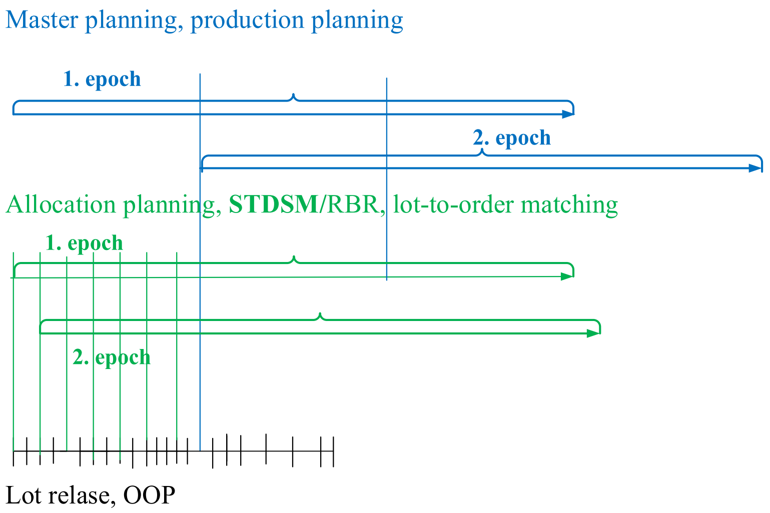

Master planning, production planning
1. epoch
2. epoch
Allocation planning, **STDSM**/RBR, lot-to-order matching
1. epoch
2. epoch
Lot relase, OOP

Fig. 5 Rolling horizon setting

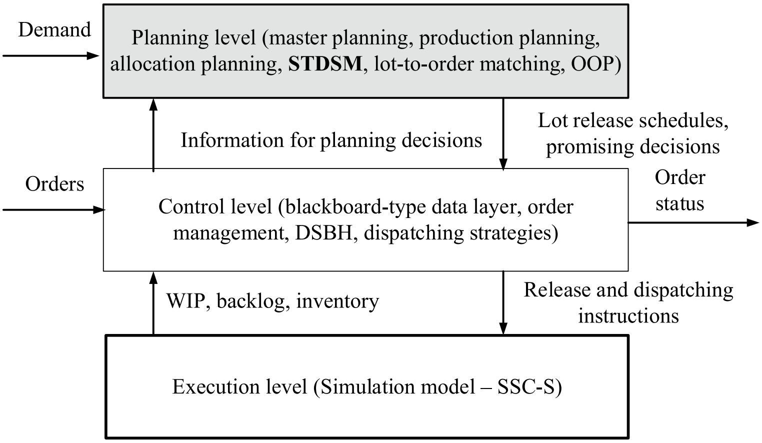

Demand
Planning level (master planning, production planning, allocation planning, **STDSM**, lot-to-order matching, OOP)
Information for planning decisions
Lot release schedules, promising decisions
Orders
Control level (blackboard-type data layer, order management, DSBH, dispatching strategies)
Order status
WIP, backlog, inventory
Release and dispatching instructions
Execution level (Simulation model – SSC-S)

Fig. 6 Simulation infrastructure

Table 2 Characteristics of the Planning and Control Process

Planning/control module	Length of the planning window	Planning frequency
Master planning	26 weeks	Weekly
STDSM	26 weeks	Daily
Allocation planning	26 weeks	Daily
Lot release planning	26 weeks	Weekly
OOP	–	Based on request
Lot-to-order matching	26 weeks	Daily
Scheduling	3 h	2 h

A slightly simplified version of the semiconductor supply chain simulation testbed (Ewen et al. 2017) that is publicly available under Testbed (2021) is used in the experiments. It contains two mature 200 mm FE facilities, each of them with more than 200 machines. Two BE facilities are also included. The semiconductor supply chain has two DBs and three DCs. Two products with more than 200 FE and 40 BE process steps are considered in the experiments. This small-sized semiconductor supply chain model is abbreviated by SSC-S. The infrastructure including the simulation model is shown in Fig. 6.

We use a simulation horizon of 2 years. It consists of $S = 728$ periods, i.e., the period length is a single day. The planning window of master planning, production planning, allocation planning, and the STDSM approach is $T = 26$ weeks. The period length is 1 week in the master and production planning approaches, while it is only a single day in allocation planning and the STDSM approach. The DSBH proposed by Mönch and Drießel (2005) serves as the scheduling approach. The next planning epoch always starts after one period. The weekly supply derived by master planning is evenly distributed over the daily planning periods of the STDSM approach. The scheduling window of the DSBH is 3 h, while a new schedule is computed every 2 h. The different lengths of the planning windows and the planning frequencies are summarized in Table 2.

4.2 Demand and order generation scheme

Demand and order information evolving over time is crucial for the rolling horizon scheme. Demand is generated based on the additive martingale model of forecast evolution (MMFE) by Heath and Jackson (1994). It is a quite general and powerful approach to model demand correlation across products and periods for production planning (Chen and Lee 2009; Albey et al. 2015; Ziarnetzky et al. 2018, 2020). The details of the MMFE demand generation scheme are provided in Appendix D of the electronic supplement for the sake of completeness.

Firm orders and forecast are differentiated in the simulation experiments. The number of firm orders decreases over time while the amount of forecast increases. Orders are needed for the OOP and the STDSM approach. Firm orders of product g at planning epoch s are modeled by $D_{st}^{(g)fo} : = D_{st}^{(g)}\left(1 - \frac{t-1}{T-1}\right)(1-\eta)$, $t = 1, \dots, T$. Here, $D_{st}^{(g)}$, $s \le t \le s + H - 1$ is the demand forecast for product g made at the end of epoch s for period t, and H is the length of the forecast window, given as number of periods. The demand of the current planning period $t = 1$ is deterministic, while the demand of future periods $t = 2, \dots, T$ consists of firm and forecasted orders that decrease and increase over time, respectively. The forecast portion is determined by the difference of total demand $D_{st}^{(g)}$ and the amount $D_{st}^{(g)fo}$ of firm orders. The bias $-1 < \eta < 0$ is used to model overestimated and $0 < \eta < 1$ to represent underestimated firm order quantities. The bias causes average realizations of orders that do not match with the forecast which is quite common in semiconductor manufacturing.

Order quantities $O_{st\tau}^{(g)(c)}$ at the planning epochs $s = 1, \dots, S$ are randomly generated according to $O_{st\tau}^{(g)(c)} \sim U\left[0.9s_{st}^{(g)(c)}, 1.1s_{st}^{(g)(c)}\right]$ for each product g, customer $c \in \Omega$, period t, and randomly chosen day $\tau = 1, \dots, 7$ of the planning period where the average order size is given by $s_{st}^{(g)(c)} = \frac{1}{7|\Omega|}\frac{\left\{D_{st}^{(g)fo} - D_{s-1,t+1}^{(g)fo}\right\}}{(1-\eta)}$. Order quantities are generated until the total amount of $\sum_{c=1}^{\Omega}\sum_{\tau=1}^{7} O_{st\tau}^{(g)(c)} = \frac{D_{st}^{(g)fo} - D_{s-1,t+1}^{(g)fo}}{1-\eta}$ is reached. The periods $7(s+t)+\tau$ represent the desired delivery date d_o of the incoming orders at planning epoch s. The demand $D_{st}^{(g)}$ of product g for each planning period is used in master planning, and it is separated into firm orders coming from order management and a forecast portion obtained as the difference between the overall demand and the firm orders. The remaining forecast portion $D_{st}^{(g)fc} : = D_{st}^{(g)} - D_{st}^{(g)fo}$ must be assigned to customers and periods with a length of a single day. The ratios of the customer-specific order quantities in already executed periods are exploited to determine the fraction of the forecast portion that is assigned to a certain customer. Customers with a large amount of orders in the past are preferred. Finally, the customer-specific forecast portion is evenly distributed over the 7 days of a master planning period.

5 Simulation experiments

We start by discussing the design of experiments in Sect. 5.1. Parameter setting issues and implementation details are described in Sect. 5.2. Section 5.3 is used to present the simulation results.

5.1 Design of experiments

We expect that the planned bottleneck utilization (BNU) of the FE facilities influences the performance of the proposed STDSM approach. Therefore, 70% and 90% are used as BNU levels of the FE facilities. Allocation planning requires a mean demand that is higher than the available ATP to model scarce capacity. A moderate and a high demand scenario are examined in the simulation experiments. The moderate demand scenario is given by BNU = 70%. The resulting demand for the moderate scenario is randomly chosen between 95 and 110% of the calculated mean demand where a continuous uniform distribution is assumed. In the high BNU scenario, the demand is randomly chosen between 110 and 130% of the determined mean demand for BNU = 90% where again a continuous uniform distribution is applied. This leads to different mean demand values μ_g of the products g for each demand scenario. Five independent demand instances are used in the experiments. Moreover, five independent simulation replications are performed for each demand instance to compute the performance measure values in the face of execution uncertainty. The average is taken over all replications. Two different cost settings are applied (see Sect. 5.2 for details). Overall, 1200 simulation runs are performed. The design of experiments is summarized in Table 3 where the abbreviation CV is used for the coefficient of variation of the demand (cf. Appendix D).

We are interested in the on-time delivery (OTD) performance defined as the fraction of all promised orders o (set O^P) that are delivered at the first promised date $\tau_o^{(i)}$ (set O^{FPD}) i.e., we have:

$$OTD: = \sum\nolimits_{o \in O^{FPD}} w_o \Big/ \sum\nolimits_{o \in O^P} w_o. \tag{19}$$

Table 3 Design of experiments

Factor	Level	Count
Planning approach	STDSM, RBR	2
Cost setting	Regular, foundry	2
BNU level	Moderate (BNU = 70%), high (BNU = 90%)	2
CV	0.10, 0.25	2
Bias η	− 0.2, 0.0, 0.2	3
Independent demand realizations		5
Independent simulation replications		5
Total simulation runs		1200

The next measure is the order-based delivery (OBD) performance:

$$OBD: = \sum\nolimits_{o\in O^D} w_o \Big/ \sum\nolimits_{o\in O^P} w_o. \tag{20}$$

It is the fraction of all promised orders that are delivered at the desired date d_o (set O^D). Moreover, the average realized waiting time (AWT) of the promised orders is measured. It is defined as:

$$AWT: = \sum\nolimits_{o\in O^P} w_o\left(\hat{\tau}_o - d_o\right)^+ \Big/ \sum\nolimits_{o\in O^T} w_o, \tag{21}$$

where the set of tardy orders (O^T) is defined by $O^T: = \left\{o \in O^P | \hat{\tau}_o - \tau_o^{(i)} > 0\right\}$. Here, $\hat{\tau}_o$ is the realized delivery date of order o. Finally, the stability is considered which is defined as

$$\phi = \frac{1}{S\sum_{o\in O^A} w_o} \sum_{o\in O^P} w_o \sum_{s=1}^{S} I(o,s)\left|\tau_o(s) - \tau_o^{(i)}\right|, \tag{22}$$

where $\tau_o(s)$ is the promised delivery date for order o in planning epoch s. Moreover, we have $I(o,s) = 1$ for the indictor $I(o,s)$ if $\hat{\tau}_o \le s$ and $I(o,s) = 0$ otherwise. A stability value close to zero is preferred since it means that the $\tau_o(s)$ value for a given order o does not change much in the different planning epochs. Investigating planning stability or instability is important for rolling horizon approaches (Kimms 1998).

5.2 Parameter setting and implementation issues

The cost settings for the STDSM approach are summarized in Table 4. Since specific cost settings are not available from the industry due to confidentiality reasons, the cost values are chosen based on our experience with semiconductor supply chains where the revenue from a single IC is much larger than inventory-related costs. Inventory holding costs are rather small relative to backlog costs since they only represent delayed revenue. FE WIP costs are chosen higher than inventory holding cost due to the limited available clean room space within wafer fabs. Additional experiments are conducted for a foundry-type setting (Mönch et al. 2018a). A pure play

Table 4 Parameter settings for the STDSM approach

FE		BE	
Parameter	Value	Parameter	Value
π_o^{DREP}	180	π_o	180
π_o^{FPOS}	150	h_{gt}^{DC}	15
h_{gt}	15 (30)	ω_{jgt}^{BE}	7
b_{gt}	50 (100)		
ω_{jgt}^{FE}	20		

foundry manufactures ICs for other companies, without designing them. This means that customer orders have to be completed on time and finished inventory cannot be used to satisfy demand of other customers. Therefore, we double backlog and inventory holding cost for the FE facilities, i.e., we use a second cost setting. The corresponding values are presented in brackets in Table 4.

The functions k, l to set the endpoints of the time window $\left[e_o, f_o\right]$ are:

$$k(r) := \begin{cases} 0 \text{ periods for } r = 1 \\ 7 \text{ periods for } r = 2 \\ 182 \text{ periods for } 3 \leq r \leq 6 \end{cases}, \quad l(s) := \begin{cases} 0 \text{ periods for } 1 \leq r \leq 3 \\ 7 \text{ periods for } r = 4 \\ 15 \text{ periods for } r = 5 \\ 182 \text{ periods for } r = 6 \end{cases}. \tag{23}$$

Note that 182 periods cover the entire planning window for the STDSM approach. The settings $\alpha_o \equiv 2$ and $\beta_o \equiv 1$ are chosen to penalize orders that are not repromised at $\tau_o^{(i)}$. We use $\alpha = 10$ periods for the decision interval and $\beta = 8$ periods for the frozen interval in the time-based decomposition approach for solving FE and BE STDSM instances. A relative MIP gap of 15% is used for each subproblem resulting from the decomposition of the MILPs. These values are chosen by trial and error based on some preliminary experimentation with a small number of problem instances. Six customers are considered in the simulation experiments. The settings $w_{cgt\tau} \equiv 5$ for the most important customer, $w_{cgt\tau} \equiv 3$ for the customer of medium importance, and $w_{cgt\tau} \equiv 1$ for the remaining four customers of regular importance are used for all products to mimic a situation where a certain customer is more important than others. The weights w_o for the resulting orders are taken as the $w_{cgt\tau}$ values of the customer. Similar to the weights of each customer, the penalty value $p_{cig\tau}$ for not reaching the minimum amount of demand that belongs to class i of product g for customer c and is due in period τ is chosen as $p_{cig\tau} \equiv 15$ for the most important customer, $p_{cig\tau} \equiv 13$ for the customer of medium importance, and $p_{cig\tau} \equiv 10$ for the four regular customers. The minimum ATP quantity $AP_{cig\tau}$ of demand class i of product g, allocated to demand for customer c which is due in period τ is set according to $AP_{c1g\tau} = \sum_{\{o|o\in A_{gc}, \tau_o = \tau\}} q_o$ and $AP_{c2g\tau} \equiv 0$. This allows prioritizing orders that are already promised as most important when allocating scarce ATP quantities. Two demand classes are used in the STDSM formulations. Demand class 1 refers to confirmed orders demand class 2 is used to model forecasted demand. The revenue and cost settings for master planning can be found in Herding and Mönch (2021).

The FE STDSM-O and FE STDSM-NO instances are solved individually for each FE facility due to the size of the FE simulation models used in the conducted simulation experiments. Therefore, we restrict the FE STDSM-O model to the orders that belong to an individual facility. Moreover, the right-hand side of constraint set (9) is replaced by $Y_{gjt}^{FE} + Y_{gjt}^{FE(i)}$ for FE facility j. For the FE STDSM-NO model, we have to replace the right-hand side of constraint set (18) by $\tilde{S}_{gjt} := y_{gjt}^{FE}$ for FE facility j where the quantities $\tilde{S}_{gjt}$ are provided by master planning. For both models, the I_{g0}^{DB} quantities are taken from master planning.

The Apparent Tardiness Cost (ATC) dispatching rule (Vepsalainen and Morton 1987) is used in the DSBH procedure. The due dates d_j of the lots for scheduling are set according to the flow factor FF, i.e. the ratio of the amount of time a lot spends in a wafer fab and the raw processing time (Mönch et al. 2013). The due date setting scheme $d_j := r_j + FF \cdot z_j \sum_{k=1}^{n_j} p_{jk}$ is used where r_j is the period when lot j is released, p_{jk} is the processing time of process step k, and n_j is the number of steps of lot j. The settings $z_j = 0.55$ and $z_j = 2.75$ are applied to make the due dates tight and wide for lots belonging to customer 1 and the remaining five customers, respectively.

All the algorithms are coded using either the C++ programming language or ILOG CPLEX 12.1. The computational experiments are performed on a computer with 3.6 GHz Intel Core i7-4790 processor with eight cores and 8 GB RAM.

5.3 Simulation results

5.3.1 Overview

The analysis of the simulation results will be centered around the following four questions which will be addressed in the subsequent subsections:

1. To which extent does the demand setting characterized by BNU and demand variability impact the results?
2. How are the results impacted by the accuracy of information for firm order quantities represented by the bias η?
3. Do the results depend on the used cost setting for the STDSM, i.e. regular vs. foundry setting?
4. What are the computing time requirements of the optimization-based and the rule-based approaches?

5.3.2 Impact of demand settings

The results of the rolling horizon experiments are shown in Table 5. 95% confidence intervals are presented instead of the values of point estimates to obtain statistically reasonable results. The results for both the RBR and the STDSM are shown. Two values are provided for each factor level and performance measure. The upper value is obtained by the RBR whereas the lower value is computed by the STDSM. Best values for each pair of performance measure values are marked bold.

Moreover, the improvement of the STDSM over the reference approach in % is shown for each factor combination in Table 6. Here, we consider the improvement value Imp: $= 100\% \left| O_{STDSM} - O_{ref} \right| \Big/ O_{ref}$, where O_{STDSM} and O_{ref} refer to the performance measure value of the STDSM and the RBR approach, respectively. Again, best results for the same factor level are marked bold. The results for the foundry setting are shown for the sake of completeness in Appendix E of the electronic supplement. Table E-1 is similar to Table 5.

Table 5 Simulation results (regular cost setting)

Factor/level		OTD	OBD	AWT (in days)	Stability
BNU					
70%	RBR	0.610 ± 0.022	0.532 ± 0.011	2.943 ± 0.039	0.682 ± 0.067
	STDSM	**0.702 ± 0.030**	**0.594 ± 0.031**	**2.663 ± 0.046**	**0.549 ± 0.054**
90%	RBR	0.579 ± 0.048	0.441 ± 0.022	4.064 ± 0.070	0.797 ± 0.042
	STDSM	**0.631 ± 0.049**	**0.541 ± 0.073**	**3.859 ± 0.087**	**0.681 ± 0.054**
CV					
0.10	RBR	0.605 ± 0.021	0.475 ± 0.015	3.144 ± 0.036	0.772 ± 0.088
	STDSM	**0.628** ± 0.021	**0.594 ± 0.045**	**3.085 ± 0.069**	**0.526** ± 0.153
0.25	RBR	0.560 ± 0.026	0.431 ± 0.021	4.179 ± 0.050	0.786 ± 0.067
	STDSM	**0.578 ± 0.036**	**0.501 ± 0.033**	**3.507 ± 0.064**	**0.607 ± 0.094**
Bias					
0.00	RBR	0.626 ± 0.015	0.473 ± 0.015	3.726 ± 0.027	0.781 ± 0.011
	STDSM	**0.670 ± 0.020**	**0.586 ± 0.021**	**3.219 ± 0.042**	**0.650 ± 0.014**
0.20	RBR	**0.545 ± 0.032**	0.353 ± 0.073	**4.225 ± 0.064**	0.922 ± 0.051
	STDSM	0.630 ± 0.027	**0.421 ± 0.018**	3.985 ± 0.073	**0.876 ± 0.027**
− 0.20	RBR	0.590 ± 0.015	0.428 ± 0.013	4.163 ± 0.040	0.821 ± 0.017
	STDSM	**0.639 ± 0.053**	**0.577 ± 0.105**	**3.596 ± 0.096**	**0.781 ± 0.075**

Table 6 Improvement of the STDSM over the RBR approach (in %) (regular cost setting)

Factor/level	OTD	OBD	AWT	Stability
BNU				
70%	**15.117**	11.732	**9.510**	**19.465**
90%	9.120	**22.825**	5.044	14.543
CV				
0.10	**3.808**	**25.174**	1.866	**31.817**
0.25	3.245	16.398	**16.066**	22.656
Bias				
0.00	7.035	24.026	**13.607**	**16.714**
0.20	**15.706**	19.425	5.664	4.952
− 0.20	8.369	**34.942**	13.605	4.775

We see from Table 5 that the STDSM approach outperforms the RBR procedure under almost all experimental conditions. Improvements of the STDSM of up to almost 35% are possible (cf. Table 6). The setting BNU = 90% leads to larger improvements for the OBD measure since it is beneficial to use the STDSM scheme in the case of scarce capacity, whereas BNU = 70% leads to larger improvements of the AWT since there is more room for improvement in this situation. The difference between the case of $CV = 0.25$ and $CV = 0.10$ is fairly small, only the AWT improvement is considerably larger for $CV = 0.25$ due to the different demand behavior. However, the improvement with respect to stability is larger for $CV = 0.10$. It turns out that the STDSM approach that simultaneously repromises all orders makes better decisions than the myopic RBR procedure. It is notable that this even holds for the stability measure.

5.3.3 Impact of information accuracy for firm order quantities

As expected, underestimated demand for firm orders, i.e. the setting $\eta = 0.2$, leads to worst performance measure values since wrong allocation and repromising decisions are made that cannot be corrected when considerably more orders arrive in the system as expected during planning. The stability is low in this situation. Overestimated demand, i.e. $\eta = -0.2$, however, leads to a situation where the performance measure values are outperformed by the ones obtained by $\eta = 0.0$ (see Table 5), but the magnitude of deterioration is smaller compared to an underestimated amount of firm orders. This behavior is caused by the fact that the number of arriving orders considered during planning is smaller. This still leads to wrong decisions, but they can be easier compensated as in the case of underestimation since the rolling horizon setting ensures periodic replanning. This is also confirmed by the results shown in Table 6 where the setting $\eta = -0.2$ consistently leads to the largest improvement rates among the different η values expect for the stability measure where $\eta = 0.0$ leads to the best results. This means that applying the STDSM scheme is particularly useful in this situation.

5.3.4 Impact of the cost setting for the STDSM scheme

The results for the foundry case are similar to the regular cost case. As expected, the OTD and OBD values in Table E-1 of the electronic supplement are slightly better than the ones for the regular case, the same is true for the AWT values since the backlog cost is much higher in the foundry setting. However, these improvements which are a result of the different cost settings are obtained at the expense of reduced stability.

5.3.5 Computing times for the optimization- and rule-based approaches

A single simulation run leads to 728 planning epochs of the STDSM approach. The corresponding average computing time for a single simulation run of the SSC-S scenario is 3581 min whereas the corresponding time for the RBR heuristic is only 1388 min. Both the STDSM and the RBR approach require allocation planning, i.e. solving instances of the model (A1)–(A6). Note that the average computing time for a single STDSM decision in the case of the SSC-S supply chain is less than 5 min. The foundry case leads to similar computing times.

6 Conclusions and future research directions

In this paper, a STDSM approach for semiconductor supply chains was proposed. The approach is based on a decomposition that takes into account the structure of the semiconductor supply chain. The NP-hardness of the related planning problem was proven. The integration of the STDSM approach into a hierarchical planning approach that includes master planning, allocation planning, and production planning was discussed. It was demonstrated by means of applying the hierarchical planning approach in a rolling horizon setting that the proposed approach outperforms a conventional rule-based repromising heuristic with respect to several delivery performance-related measures and with respect to stability.

Next, we provide managerial insights for decision makers in the semiconductor industry:

1. Optimization-based approaches for STDSM are able to outperform more conventional rule-based approaches with respect to on-time delivery-related performance measures. This is especially true in situations where the capacity in the FE facilities is scarce.
2. Moreover, optimization-based approaches are able to increase the planning stability to a large extent, especially when the demand uncertainty is fairly small. Nervous plans are not desirable since customers accept only a few changes in the promised delivery date.
3. The accuracy of estimating firm order quantities is crucial for the performance of the optimization-based approaches. However, underestimation is more critical than overestimation.
4. Although the planning problems are NP-hard, the computing times for the optimization-based approaches are acceptable if heuristic decomposition approaches are applied. They can be further improved if the cloud-based infrastructure proposed by Herding and Mönch (2022) is used or if metaheuristic approaches are applied (Wang and Mönch 2021).
5. The proposed optimization-based approach seems to be quite implementable in next-generation software agent-based planning systems for semiconductor supply chains (Herding and Mönch 2022).

There are several directions for future research. First of all, we believe that it is possible to execute the STDSM approach in a distributed manner using a cloud-based infrastructure to obtain reasonable computing times. Cloud manufacturing is a promising direction for semiconductor supply chains (Wu et al. 2014; Chen 2014; Yang et al. 2020; Herding and Mönch 2022). As a second direction of future research, we propose using the simulation-based infrastructure designed by Herding and Mönch (2022) to include more directly demand and execution uncertainty in procedures for STDSM via simulation-based optimization.

As a third research avenue, it seems possible to replace the time-based decomposition approach to solve the FE STDSM-O and BE STDSM models by metaheuristic approaches. The meta-heuristic chooses the values of the binary decision variables associated with the repromising decisions while an LP approach deals with the remaining decisions. The approach will require solving many LPs. Such models are already proposed by Wang and Mönch (2021) for single wafer fabs. The approach could take advantage of the proposed distributed computing environment. Moreover, it is well-known that the planning models based on exogenous lead times are outperformed by models with workload-dependent lead times (Mönch et al. 2018b). Therefore, formulations based on non-linear clearing functions have to be investigated as a fourth research direction.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10100-023-00891-z.

Acknowledgements The authors would like to thank Alexander Seitz, Thomas Ponsignon, and Hans Ehm, Infineon Technologies AG, Germany for fruitful discussions on demand fulfillment in semiconductor supply chains.

Author contributions RH: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, writing-original draft, visualization, validation, writing-reviewing and editing. LM: Conceptualization, methodology, validation, formal analysis, investigation, resources, writing-original draft, visualization, validation, writing-reviewing and editing.

Funding Open Access funding enabled and organized by Projekt DEAL.

Availability of data and material The simulation models used in the experiments are publicly available under https://p2schedgen.fernuni-hagen.de/downloads/simulation. Order instances and detailed computational results can be provided based on request.

Code availability The source code is not publicly available.

Declarations

Conflict of interest There are no conflict of interest.

References

Albey E, Norouzi A, Kempf K, Uzsoy R (2015) Demand modeling with forecast evolution: an application to production planning. IEEE Trans Semicond Manuf 28(3):374–384

Alemany MME, Ortiz A, Fuertes-Miquel VS (2018) A decision support tool for the order promising process with product homogeneity requirements in hybrid make-to-stock and make-to-order environments. Applications to a ceramic tile company. Comput Ind Eng 122:219–234

Azevedo RC, Amours SD, Rönnqvist M (2016) Advances in profit-driven order promising for make-to-stock environments—a case study with a Canadian softwood lumber manufacturer. INFOR Inf Syst Oper Res 54(3):210–233

Babarogić S, Makajić-Nikolić D, Lečić-Cvetković D, Atanasov N (2012) Multi-period customer service level maximization under limited production capacity. Int J Comput Commun Control 7(5):798–806

Brahimi N, Aouam T, Aghezzaf EH (2015) Integrating order acceptance decisions with flexible due dates in a production planning model with load-dependent lead times. Int J Prod Res 53:3810–3822

Chen T (2014) Strengthening the competitiveness and sustainability of a semiconductor manufacturer with cloud manufacturing. Sustainability 6:251–266

Chen J, Dong M (2014) Available-to-promise-based flexible order allocation in ATO supply chains. Int J Prod Res 52(22):6717–6738

Chen L, Lee H (2009) Information sharing and order variability control under a generalized demand model. Manag Sci 55(5):781–797

Chen JH, Lin JT, Wu YS (2008) Order promising rolling planning with ATP/CTP reallocation mechanism. Ind Eng Manag Syst 7(1):57–65

Chiang C, Hsu H-L (2014) An order fulfillment model with periodic review mechanism in semiconductor foundry plants. IEEE Trans Semicond Manuf 27(4):489–500

Chien C-F, Dauzère-Pérès S, Ehm H, Fowler J, Jiang Z, Krishnaswamy S, Mönch L, Uzsoy R (2011) Modeling and analysis of semiconductor manufacturing in a shrinking world: challenges and successes. Eur J Ind Eng 5(3):254–271

Chien C-F, Ehm H, Fowler JW, Mönch L (2016) Modeling and analysis of semiconductor supply chains (Dagstuhl seminar 16062). Dagstuhl Rep 6(2):28–64

Dowsland KA, Dowsland WB (1992) Packing problems. Eur J Oper Res 56(1):2–14

Ewen H, Mönch L, Ehm H, Ponsignon T, Fowler JW, Forstner L (2017) A testbed for simulating semiconductor supply chains. IEEE Trans Semicond Manuf 30(3):293–305

Fleischmann B, Meyr H (2004) Customer orientation in advanced planning systems. In: Dyckhoff H, Lackes R, Reese J, Fandel G (eds) Supply chain management and reverse logistics. Springer, Berlin, pp 298–321

Fordyce K, Wang C-T, Chang C-H, Degbotse A, Denton B, Lyon P, Milne RJ, Orzell R, Rice R, Waite J (2011) The ongoing challenge: creating an enterprise-wide detailed supply chain plan for semiconductor and package operations. In: Kempf K, Keskinocak P, Uzsoy R (eds) Planning production and inventories in the extended enterprise: a state of the art handbook, vol 2. Springer, Berlin, pp 313–387

Framinan JM, Perez-Gonzalez P (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, pp 2653–2663

Geier S (2014) Demand fulfillment bei assembly-to-order-fertigung: analyse, und anwendung in der computerindustrie. Springer, Wiesbaden

Heath DC, Jackson PL (1994) Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. IIE Trans 26:17–30

Herding R, Mönch L (2021) A short-term demand supply matching approach for semiconductor supply chains. Technical report (INFORMATIK BERICHTE 382–06/2021), University of Hagen, Department of Mathematics and Computer Science

Herding R, Mönch L (2022) An agent-based infrastructure for assessing the performance of planning approaches for semiconductor supply chains. Expert Syst Appl 202:117001

Herding R, Mönch L, Ziarnetzky T, Ponsignon T, Seitz A, Ehm H (2017) Simulation-based performance assessment of demand fulfillment approaches in semiconductor supply chains. In: Proceedings of the 47th international conference on computers and industrial engineering (CIE47)

Jeong B, Sim SB, Jeong HS, Kim SW (2002) An available-to-promise system for TFT-LCD manufacturing in supply chain. Comput Ind Eng 43:191–212

Kacar NB, Mönch L, Uzsoy R (2016) Modeling cycle times in production planning models for wafer fabrication. IEEE Trans Semicond Manuf 29(2):153–167

Kallrath J, Maindl TI (2006) Planning in semiconductor manufacturing. In: Kallrath J, Maindl TI (eds) Real optimization with SAP APO. Springer, Heidelberg, pp 105–117

Kilger C, Meyr H (2015) Demand fulfilment and ATP. In: Stadtler H, Kilger C, Meyr H (eds) Supply chain management and advanced planning—concepts, models and software, 5th edn. Springer, Berlin, pp 177–194

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

Knutson K, Kempf K, Fowler JW, Carlyle M (1998) Lot-to-order matching for a semiconductor assembly and test facility. IIE Trans 31:1103–1111

Leachman RC, Benson F, Liu C, Raar DJ (1996) IMPReSS: an automated production-planning and delivery-quotation system at Harris corporation: semiconductor sector. Interfaces 26(1):6–37

Lin J, Hong I, Wu C, Wang K (2010) A model for batch available-to-promise in order fulfillment processes for TFT-LCD production chains. Comput Ind Eng 59(4):720–729

Lyon P, Milne R, Orzell R, Rice R (2001) matching assets with demand in supply-chain management at IBM microelectronics. Interfaces 31(1):108–124

Meyr H (2009) Customer segmentation, allocation planning and order promising in make-to-stock production. Or Spectrum 31:229–256

Missbauer H, Uzsoy R (2020) Production planning with capacitated resources and congestion. Springer, New York

Missbauer H, Uzsoy R (2022) Order release in production planning and control systems: challenges and opportunities. Int J Prod Res 60(1):256–276

Mönch L, Drießel R (2005) A distributed shifting bottleneck heuristic for complex job shops. Comput Ind Eng 49(3):673–680

Mönch L, Fowler JW, Mason SJ (2013) Production planning and control for semiconductor wafer fabrication facilities: modeling, analysis and systems. Springer, New York, p 52

Mönch L, Uzsoy R, Fowler JW (2018a) A survey of semiconductor supply chain models part I: semiconductor supply chains, strategic network design, and supply chain simulation. Int J Prod Res 56(13):4524–4545

Mönch L, Uzsoy R, Fowler JW (2018b) A survey of semiconductor supply chain models part III: master planning, production planning, and demand fulfillment. Int J Prod Res 56(13):4565–4584

Mousavi BA, Azzouz R, Heavey C (2019) Mathematical modelling of products allocation to customers for semiconductor supply chain. Procedia Manuf 38:1042–1049

Ponsignon T, Mönch L (2012) Heuristic approaches for master planning in semiconductor manufacturing. Comput Oper Res 39(3):479–491

Seitz A, Grunow M (2017) Increasing accuracy and robustness of order promises. Int J Prod Res 55(3):656–670

Seitz A, Grunow M, Akkerman R (2020) Data driven supply allocation to individual customers considering forecast bias. Int J Prod Econ 227:107683

Smith KD, Kempf KG (2005) Application of model predictive control and optimization methods to semiconductor manufacturing supply-side inventory replenishment. In: Proceedings of the IEEE international symposium on semiconductor manufacturing, pp 35–38

Soares A, Azevedo A, de Sousa J (2000) Distributed planning and control systems for the virtual enterprise: organizational requirements and development life-cycle. J Intell Manuf 11:253–270

Testbed (2021) https://p2schedgen.fernuni-hagen.de/downloads/simulation. Last Accessed 26 Sept 2021

Tsai K, Wang S (2009) Multi-site available to promise modeling for assemble-to-order manufacturing: an illustration on TFT-LCD manufacturing. Int J Prod Econ 117:174–184

Vepsalainen APJ, Morton TE (1987) Priority rules for job shops with weighted tardiness cost. Manag Sci 33(8):1035–1047

Wang H-K, Mönch L (2021) A matheuristic for making order acceptance decisions in multi-product, multi-stage manufacturing systems. Appl Soft Comput 211:107640

Wu X, Qiao F, Poon K (2014) Cloud manufacturing application in semiconductor industry. In: Proceedings of the 2014 winter simulation conference, pp 2376–2383

Yang H-C, Chen Y-L, Cheng F-T (2020) A cloud manufacturing service to control outsourcing production. J Chin Inst Eng 43(8):838–850

Zhao Z, Ball M, Chen CY (2003) A scalable supply chain infrastructures research test-bed. In: Prabhu V, Kumara S, Kamath M (eds) Scalable enterprise systems: an introduction to recent advances. Kluwer, Boston, pp 217–247

Ziarnetzky T, Mönch L, Uzsoy R (2018) Rolling horizon, multi-product production planning with chance constraints and forecast evolution for wafer fabs. Int J Prod Res 56(18):6112–6134

Ziarnetzky T, Mönch L, Uzsoy R (2020) Simulation-based performance assessment of production planning models with safety stock and forecast evolution in semiconductor wafer fabrication. IEEE Trans Semicond Manuf 33(1):1–12

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.