Optimizing the workload of production units of a make-to-order manufacturing system

Murat Firat\textsuperscript{a,}\textsuperscript{*}, Julie De Meyere\textsuperscript{c}, Tugce Martagan\textsuperscript{b}, Laura Genga\textsuperscript{b}

\textsuperscript{a} Department of Computer Science, Open University, The Netherlands
\textsuperscript{b} School of Industrial Engineering, Eindhoven University of Technology, The Netherlands
\textsuperscript{c} Applied Micro Electronics AME B.V., The Netherlands

\begin{abstract}
This paper proposes a production planning approach for a job shop type manufacturing company in electronics sector that operates with make-to-order (MTO) convention and has a high-mix production range. A novel Mixed Integer Linear Programming (MILP) model is proposed that finds workload-dependent planning horizon by making order acceptance decisions. Partial acceptance of orders and delayed internal target deliveries contribute to the novelty of the proposed model. We establish the complexity of the problem and solve real-life problem instances of a medium-size manufacturing company in the south of The Netherlands. Our computational results show that the proposed MILP model can allocate desired workload to work centers by dynamically adjusting the time horizon for production plans. Our approach can practically be used within reasonable times for daily updates of internal production targets. We perform a sensitivity analysis to show how the time horizon is determined with varying system conditions. The obtained workload amounts satisfy minimum requirements of production units, as this cannot be achieved by the current practice.
\end{abstract}

\section{Introduction}

Efficient production planning and capacity management is a crucial as well as challenging task in MTO (make-to-order) flexible job-shop type production companies with high-mix product range. We study in this paper a multi-stage production system that has the following characteristics. The fixed production paths of final products specify the work centers to visit for different types of tasks, each can be executed in alternative machines. Products have highly customized design and therefore they have customer-specific product numbers, hence the product range is high-mix. The efficiency in the considered production setting requires a high degree of interdependence between separate production units within the production system. Coordination between multiple production units in a multi-stage production structure is a challenge as not only the timing of the production needs to be considered, but also the capacity of the production resources for each component that makes up the final product (Jodlbauer and Strasser, 2019; Toczyłowski et al., 1989).

In Industry 4.0 era, the companies are required to have high responsiveness to changing, hence challenging, market conditions (Lu et al., 2019). Especially, detecting as well as reacting to unplanned situations, i.e. late supplier deliveries, process quality drops, worker unavailability and so on, in the best way is of high importance. In this paper, we develop an approach considering sales and customer-confirmed forecasted order information with deadlines instead of only forecasted demand as in traditional Manufacturing Planning and Control (MPC) procedures. Therefore, the output of our approach includes actionable production targets so that various ways to determine operational decisions can be taken from conventional production scheduling to agent-based online job shop scheduling under advanced capabilities of Industry 4.0 (Leusin et al., 2018).

The studied production system has a size close to SME (small and medium-sized enterprises) definition, therefore our approach is planned to be a enabler, as a connecting part of SME-compatible information technologies for achieving potential advancements under Industry 4.0 (Rogalski et al., 2013).

Recently, there are pioneering studies tackling integrated process planning and scheduling problem, however they can solve limited instance sizes, e.g. up to 25 jobs (Sobeyko and Mönch, 2017) and 80 jobs (Shokouhi, 2018). This indicates the need for large-scale optimization approaches.

In conventional MPC, order selection, capacity management and production planning decisions tend to be treated as separate decision processes, and Master Production Schedule (MPS) is generally defined...
at the level of final products if the product diversity is relatively low (Zijm and Regattieri, 2019). The production planning is left to the production department, while the order acceptance is left to the sales department. Accepting unrealistic orders, too many or too little, may cause an inefficient utilization of the production system, which in turn will result in decreased customer satisfaction due to longer lead times or under-utilization of production units (Ebben et al., 2005).

In this work, we solve a short-term process planning problem with confirmed customer orders. For each production unit, a minimum workload is required and the planning horizon should accordingly include corresponding orders. Some orders may be rejected due to capacity constraints of the production system, and the accepted orders may have some delays in their target deliveries. Fig. 1 shows the workload-dependent planning horizon selection for two order sets with high density and low density. In the low density case, the planning horizon is selected with a longer length. However, this leads to some order rejections for work centers that are highly loaded. The advanced control and monitoring capability of our approach brings high competitive power to company management. Our production planning approach contributes to the integration of order management, production scheduling, and material requirement, hence it is a promising component for internal flexibility of a production control of the considered manufacturing system.

Our contributions. The contribution of this paper is fourfold. First, we analyze a real-life production planning problem of a medium-size enterprise in electronics sector of the Netherlands. Second, our workload optimization model is capable of handling order management decisions with possible delayed delivery target dates as well as determining internal production targets by respecting production system capacity. To our knowledge, no production planning model found in literature used a workload-dependent time horizon for order selection. Third, the selected planning horizon is so-called strong by considering partial production requirements of (confirmed) orders that have close deadlines to the planning horizon (Chung et al., 1988). Last, the experimentation results show that near-optimal solutions to real-life problem instances can be obtained within two hours of computation time, making the approach suitable to be used on a daily basis.

This paper is organized as follows. In Section 2, we review literature that has been done on comparable topics. Section 3 describes the production application studied in this paper and the research objective. Section 4 gives the basic information about the production system of the studied manufacturing company. The planning problem that we tackle is introduced as well in this section. Section 5 presents the proposed planning approach and our MILP formulation. Section 6 includes the computational experiments and sensitivity analysis of the proposed method. Lastly, our conclusions and further research directions are given in Section 7.

2. Related work

The proposed production planning approach in this paper is a short-term method according to the Supply Chain Planning Matrix (Fleischmann et al., 2015). It is the first part of a two-step production control approach and followed by a scheduling module. The literature is limited on the capacity planning methods with order acceptance aspect. Due to minimizing the rejected orders our approach can be categorized as order rejection method in the classification of Slotnick (2011).

If order selection decision is not connected to production plan information, there may either be too many orders for the capacity of the system, or an under-utilization of the production capacity. Models in literature however have different approaches on how they define the order acceptance function, as will be described below.

2.1. Order acceptance/rejection

The work by Ebben et al. (2005) studies workload based order acceptance in job shop environments. Order arrivals are stochastic and five production resources are considered. The authors use simulation with priority rules for production scheduling of accepted orders. An integrated order acceptance and production scheduling topic is studied by Chen et al. (2009) for a MTO type job shop system with three resources. The authors proposes a MILP model and solve a hypothetical problem instance of four orders, each requires four operations. The model proposed in this paper considers the short-term capacity planning problem as opposed to the medium-term capacity planning problem, however uses a similar workload-based acceptance mechanism as proposed in the model in this paper. The reason as to why the model in this paper does not directly convert the orders into a schedule is due to significantly more complex production system than the one considered by Chen et al. (2009).

A recent study by Aouam et al. (2018) develops a MILP-based approach to order acceptance and production planning problem considering uncertainty in order quantities. The authors propose two-step heuristic for solving the presented MILP model. The main idea of their heuristic is to solve the relaxed model by keeping a subset of decision variables as integer. Then in the next iteration the optimal values of the integer variables are fixed and another subset of variables are defined as integer.

A common practice is to accept all orders that enter the production system, regardless of whether they are feasible given the production capacity or not. According to Philipoom and Fry (1992), it may sometimes be better to reject specific orders than accept it and deliver it late. Accepting it may strain the workload of the work centers, and may hereby also threaten the timely delivery of other orders. The acceptance decision is made simultaneously with the assignment of a due date. Moreira and Alves (1992) describes three rules which can be used to accept or reject customer orders: total acceptance (TA), acceptance based on the present and future workload (PFW), presented by Nandi (2000), and due date negotiation (DDN) presented by Moreira and Alves (2006). In TA all orders are accepted automatically. In the PFW rules the arriving orders are accepted if they do not cause for the workload limit to be exceeded, otherwise they are rejected.

The DDN mechanism, allows for renegotiation of the order delivery date (i.e. delivery delay) if the workload exceeds the pre-defined workload limit. This paper however does not deal with a multi-level and multi-work center production structure, and leaves the order delay or rejection up to negotiations with the customer, as opposed to incorporating these functions within the model.

To our knowledge, all studies in the literature consider production systems with few resources, hence having a certain gap between real-life production systems. This paper makes a pioneering attempt to consider a production system of five production units, each having several alternative machines for the operations. Moreover, the order management methods in the literature generally either accept or reject orders. Our approach also decides deadline adjustments of accepted orders with respect to the production system capacity.

2.2. Planning horizon

The planning horizon is the length of time in the future for which the planner develops activity schedules (Chung et al., 1988). If the optimal schedule of a planning horizon is completely independent of the demand information in the future, then the planning horizon is said to be strong. Our approach finds always strong planning horizons due to considering the orders partially if some of their lead time stays within the planning horizon.

Planning horizon including approaches developed in literature are mostly based on single-product problems. The difficulty in determining
the planning horizon in multi-item models is because of complex capacity constraints (Baker, 1977). One of the first papers studying planning horizon, published by Chung et al. (1989), proposes a heuristic for finding the planning horizon for hierarchical production planning. In this heuristic a planning horizon is determined in order to smoothen the production with minimizing the inventory costs as the main objective.

Kunreuther and Morton (1973) have developed a method of determining the planning horizon that they refer to as the “natural planning horizon” as it places the planning horizon at the point where the optimal solution is the same as the optimal solution of all subsequent periods. Another method, that was proposed by Chung (2000), so-called “rolling planning horizon”, determines the planning horizon by iterative updates through solving the aggregate planning model, implementing the decisions in the scheduling, updating the model, and re-solving the problem. The actual implementation of the aggregate production planning horizon on the master schedule will provide feedback for the next planning session.

2.3. Capacity management

Recently Aouam et al. (2018) studied the single-product order acceptance in MTO environment problem by considering the production levels, production setups, and inventory levels under uncertain demand conditions. The authors formulated the production planning as a MILP model that assumes the product is produced in one production unit. A decision making structure is developed for managing arriving orders using firstly order prioritization and secondly rough-cut capacity planning to reject or accept orders.

A recent work by Aouam et al. (2018) proposes a more broadly-applicable model that integrates capacity lot-sizing along with order acceptance. The orders are rejected if their acceptance result in additional production setups with small production lots. However, delaying the deadlines is not considered. Although these studies use workload requirements as a base of order acceptance, they rely heavily on a prioritization of orders based on factors such as setup times, revenues and costs. Ponsignon and Mönch (2012) considers both (confirmed) customer orders and demand forecasts. All confirmed orders are to be produced, but forecasted demand amounts can be partially produced. Ebben et al. (2005) proposes several different approaches that can be used to integrate order acceptance with resource capacity loading, and shows that integrating these two functions can significantly improve the performance of the system, especially in the case of tight due dates.

2.4. Workload control

Workload control is a production planning and control approach that has been designed to meet the needs of complex production systems such as make-to-order and job-shop style companies. The main principle of it is to balance the input (the orders) with the capacity available within the system (the capacity of the work centers). It is also referred to as an Input–Output Control method where the goal is to level the input (i.e. the orders) with the output (i.e. the products output rate based on the capacity of the work center) (Stevenson and Hendry, 2006).

Workload control is highly beneficial in MTO environments in order to ensure that delivery dates can be met. Fattahi and Khodadad (2015) proposes a hierarchical method with four different levels of planning based on work load control. These levels are (1) enquiry stage, acceptance or rejection of orders, (2) (re-)negotiation of delivery time and price as well as acquiring the raw materials, (3) the accepted orders are released onto the shop floor, and (4) the last stage prioritization of the orders within workstations.

To sum up, the literature review in this section shows the lack of an approach covering all aspects in our proposed model; short-term production planning, a workload-dependent time horizon, workload based order acceptance, and combined order deadline determination.

3. Application background

We study the production planning problem of an electronics manufacturing company Applied Microelectronics (AME) B.V., located in Eindhoven, the Netherlands. In this job shop production setting, final products require multi-stage production with fixed processing routes. Every product is customer-specific and production is conducted in MTO convention. There are four main production units, i.e. work centers, each producing disjoint set of components. Fig. 2 illustrates the work centers of the considered job shop production system with arcs that represent the frequency of used process routes in year 2019.

In the MTO setting, an accepted customer order defines production targets for each system resource required by the corresponding process route. One example of process routes is given in Fig. 3. As it is seen, every component in the production steps of a final product has a product number, shortly PN, 12 digits. Having a PN number for every component is aslo useful for the case that a component for one order may be a final product for another order. Labeling all components with
a product number is aligned, even required, with the digitalization at the company.

The challenge of the considered production problem is having high number of customized products that requires simultaneous handling of different process routings. Hence advanced control of production capacity by structured planning is crucial to minimize delivery delays. This will eventually decrease the risk of reduced customer satisfaction.

Fig. 4 shows seven resources in the production system of AME. Note that some workcenters, e.g. Injection Molding (IM), have several resources due to different technologies of production.

4. Problem description

4.1. Basic notation

This section introduces basic concepts and the corresponding notation that is used in our MILP formulation. The problem instance has three sets of objects; product types, production units, and customer orders. All planning decisions are to be made in a set of time units \( T \) that is indexed by \( t \).

**Production units.** The production system contains a set \( R \) of resources. A resource \( r \in R \) has production capacity \( U_{r,t} \) within a time period of length \( t \). In order to limit idle times of production units, especially the expensive ones, a minimum workload is defined for every resource and it is denoted by \( H_r \), for all \( r \in R \). In other words, \( H_r \) is the (minimum) time length that the production unit \( r \) is expected to be non-idle in the planning horizon. Minimum workload requirement plays an important role to determine the time horizon of the planning for a given set of orders.

**Orders.** Each order in set \( O \) represents a demand of one final product type \( i \in P \) requested by one customer. Production routing of a final product includes the required operations associated with corresponding precedence relations. Each order \( o \in O \) has a quantity \( q_o \) and deadline \( d_o \). We define latest production start date \( l_{s_o} \), so that any planning horizon including the time window \([l_{s_o},d_o] \), even partially, should make an acceptance or rejection decision for the order \( o \). If the time window partially falls into planning horizon and the order is accepted, then the order quantity is partially considered with respect to the intersection ratio. This gives the independence property of our planning horizons. For order \( o \), the remaining processing time after production of the necessary amount of item \( i \) is denoted by \( F_{o,i} \).

**Planning horizon.** The decision variable \( r \) models the planning horizon to select the (non-rejected) orders for production targets. Accordingly, selection of orders and their contribution to production targets are determined via use of several decision variables as listed and described in Table 1. The binary variable \( \theta_{d_o} \) indicates the selected target delivery for order \( o \) at time \( t \), possibly later then the deadline \( d_o \). If the capacity of the system is not enough to produce order \( o \) within its tolerance delay period, then it gets rejected that is indicated by binary variable \( r_{o,d} \). The ratio of the interval \([l_{s_o},d_o] \) that stays within the time horizon determines the portion of order amount \( q_o \) to produce that is the value of decision variable \( r_{o,d} \). Binary variables \( x_i \) and \( y_i \) are used to determine the production ratio regarding the time horizon. The relation of these variables are case-wise shown in Fig. 5.

Each order with \( p_o > 0 \) in the solution is included within the production targets. Target delivery data of an order can be delayed, with respect to its deadline, for a time not longer than \( C \) that is called customer order delay tolerance. The deadline of order \( o \) can be delayed between \([d_o,d_o + C] \) that is indicated by one of the \( \theta \) variables. If the order needs to be delayed more than \( C \) then the order must be rejected \((r_o = 1) \). For each product type \( i \) the required inventory quantity at time \( t \) is found by decision variables \( I_{o,i} \). The number of setups necessary for component \( i \) till time \( t \) is modeled by decision variable \( \psi_{o,i} \).

5. Solution approach

In hierarchical production planning, production and operations planning are often considered as a multi-level sequential process where the decisions made in each level are used as input for the subsequent lower level (Leusink et al., 2018). Our planning approach is the first step of a decomposition of the entire manufacturing process into two sequential decision models: (1) planning of internal production targets and (2) constructing the production schedule. The proposed method in this work delivers a plan of internal production targets by solving a MILP model. Our MILP model is generic to solve the planning problems of all MTO job shop type production systems. Production targets are the quantities of each component type that need to be produced in every time unit.
It is formulated as in the following expression.

The complete sets, parameters and decision variables are summarized in Table 1. The objective (1) of the model has three components. Primary objective is to minimize the rejected orders. Secondary objective includes the planning horizon and total delay amount of deadlines. It is formulated as in the following expression.

Constraints (2) and (3) relate the planning horizon to the consideration cases of an order in production targets. These cases are shown in Fig. 5 with corresponding values of decision variables and r values.

Constraints (4) ensure that at most one of rejection or contribution to the production targets can happen.

Constraints (5) ensure that if an order is rejected, then the time horizon should be smaller than the time where order o can be maximally delayed.

Constraint pair (6) set the bound for the number items in production targets due to order o. Non-rejected orders with y = 1 are allowed to contribution to production targets. Orders with y = 1 contribute to production targets with their full quantities. Constraints (7) ensure that when an order only partially falls within the planning horizon, the order quantity that contributes to the production targets ρ o is proportionally equal to the length of the active time that falls within the planning horizon.

Fig. 4. The resources of the workcenters considered in our case study of AME.

Table 1
Sets, Parameters, Decision variables.

<table>
<thead>
<tr>
<th>Sets</th>
<th>Parameters</th>
<th>Decision variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>Set of all customer orders, indexed by o.</td>
<td>Planning horizon.</td>
</tr>
<tr>
<td>P</td>
<td>Set of all components and products, indexed by i.</td>
<td>ρ o, Number of items to produce for order o.</td>
</tr>
<tr>
<td>M</td>
<td>Set of all resources, indexed by r.</td>
<td>ρ r, Binary indicating that order o is accepted.</td>
</tr>
<tr>
<td>T</td>
<td>Set of all time units between t = {0, 1, ..., T}, indexed by t.</td>
<td>x o, Binary indicating that order o is accepted with complete quantity, not partial.</td>
</tr>
<tr>
<td>P r</td>
<td>Set of components required for final product of order o. P r ⊂ P</td>
<td>y o, Binary indicating that order o is accepted.</td>
</tr>
<tr>
<td>P r</td>
<td>Set of components produced in production resource r. P r ⊂ P</td>
<td>r o, Binary indicating that order o is rejected.</td>
</tr>
<tr>
<td>M</td>
<td>Big-M parameter.</td>
<td>a l s, Number of component i to be produced to produce one product for order o.</td>
</tr>
<tr>
<td>l s</td>
<td>Latest production start / deadline of order o.</td>
<td>H r, Time length of minimum workload required for resource r.</td>
</tr>
<tr>
<td>p i</td>
<td>Processing time of component i.</td>
<td>U r, Utilization ratio of the resource r.</td>
</tr>
<tr>
<td>a i</td>
<td>Number of component i to be produced to produce one product for order o.</td>
<td>s i, Setup time required for producing component i.</td>
</tr>
<tr>
<td>H r</td>
<td>Time length of minimum workload required for resource r.</td>
<td>C r, Customer order delay tolerance.</td>
</tr>
<tr>
<td>U r</td>
<td>Utilization ratio of the resource r.</td>
<td>F r, i, Remaining processing time of order o after necessary number of component i is produced.</td>
</tr>
<tr>
<td>s i</td>
<td>Setup time required for producing component i.</td>
<td>Ψ i, The number of setups necessary for producing component i by time t.</td>
</tr>
<tr>
<td>C r</td>
<td>Customer order delay tolerance.</td>
<td>R r, Batch size of component i.</td>
</tr>
<tr>
<td>F r, i</td>
<td>Remaining processing time of order o after necessary number of component i is produced.</td>
<td></td>
</tr>
</tbody>
</table>

The proposed MILP model simultaneously makes order acceptance/rejection decisions and converts accepted orders to production targets. The model respects the utilization ratios and desired amount of workload for high-tech production units. Customer orders, processing routes, and production system properties are the input for the model construction. The main objective of the model is minimizing the rejections of orders. Target delivery dates may be later than the deadlines provided that they stay within customer delay tolerance.

The selected planning horizon and accepted orders must result in feasible, also efficient, workload of production units. Expensive production units should have enough workload within a pre-determined time period within the planning horizon. Accepted customer orders are converted to the workload of production units in discrete amounts with corresponding deadlines by considering the processing times of operations to be done.

5.1. MILP formulation

It is assumed that processing times are deterministic, machine breakdowns are negligible, and production units can be used in a continuous time line. The production of every product is assumed in the same amounts of the batch size.

The complete sets, parameters and decision variables are summarized in Table 1. The objective (1) of the model has three components. Primary objective is to minimize the rejected orders. Secondary objective includes the planning horizon and total delay amount of deadlines. It is formulated as in the following expression.

Minimize \( \frac{r}{T} + \sum_{o \in O} \left( M r_o + \sum_{r \in P_r} x o \right) \) \( \sum_{i \in I} \left( t - d i - a i \right) \) \( \rho r \).

Constraints (2) and (3) relate the planning horizon to the consideration cases of an order in production targets. These cases are shown in Fig. 5 with corresponding values of decision variables and r values.

Constraints (4) ensure that at most one of rejection or contribution to the production targets can happen.

Constraints (5) ensure that if an order is rejected, then the time horizon should be smaller than the time where order o can be maximally delayed.

Constraint pair (6) set the bound for the number items in production targets due to order o. Non-rejected orders with y = 1 are allowed to contribution to production targets. Orders with y = 1 contribute to production targets with their full quantities. Constraints (7) ensure that when an order only partially falls within the planning horizon, the order quantity that contributes to the production targets ρ o is proportionally equal to the length of the active time that falls within the planning horizon.
By constraints (8) every non-rejected order has one target delivery date. Constraints (9) ensure that a non-rejected order with full quantity contribution has target delivery date between \( d_o \) and \( d_o + C \).

\[
\sum_{i=1}^{d_o+C} \theta_{o,i} = x_o, \quad \forall o \in O
\]

(8)

\[
\sum_{i=1}^{d_o+C} \theta_{o,i} \geq y_o, \quad \forall o \in O
\]

(9)

Constraints (10) ensure that target delivery times of selected orders stay within the planning horizon, i.e. not larger than \( T \) value. A partially included order \( o \) has \( \theta_{o,i} = 1 \) (due to \( x_o = 1, y_o = 0 \)). In rejection case \( (r_o = 1) \), both constraints become redundant (due to \( x_o = y_o = 0 \)). The target delivery of an order will contribute to production level of its corresponding components in constraints (11). Production levels are also subjected to resource capacities by constraints (12).

\[
M(x_o - y_o - 1) + r \preceq \sum_{i=1}^{d_o} \theta_{o,i}, \quad \forall o \in O.
\]

(10)

Constraints (11) ensure that accepted orders contribute to production targets for the required product types at corresponding times with respect to their selected target delivery dates. The contributions of accepted orders are determined by going downstream production levels in their processing routes.

\[
I_{o,r} - I_{o,r-1} = \sum_{i=1}^{d_o+C} \left( \min \left( 1, \frac{t - I_{o,i}}{d_o - I_{o,i}} \right) \right) q_o \theta_{o,i}, \quad \forall r \in T, \quad i \in P.
\]

(11)

Constraints (12) ensure that the workload assigned to the resource \( r \) can be produced in each time unit. Constraints (13) ensure that the minimum workload amount is collected within the selected planning horizon.

\[
\sum_{i \in T} \left( s_i \psi_{i,r} + p_i I_{i,r} \right) \leq t U_r, \quad \forall r \in R, \quad i \in T.
\]

(12)

\[
\sum_{i \in T} \left( s_i \psi_{i,r} + p_i I_{i,r} \right) \geq H_i U_r, \quad \forall r \in R.
\]

(13)

Constraints (14) set the number of setups necessary by time \( t \) for the batch sizes of product types.

\[
I_{o,i} \leq B_{o,i}, \quad \forall r \in T, \quad i \in P.
\]

(14)

Constraints (15) and (16) specify the bounds of the decision variables.

\[
x_o, y_o, \theta_{o,i} \in \{0, 1\}, \quad \forall r \in T, \quad o \in O
\]

(15)

\[
I_{o,i}, \phi_{i,j} \geq 0, \quad \forall r \in T, \quad i \in P
\]

(16)

Model size. The size of the MILP model (2)–(14) has two main groups of constraints; order related constraints (2)–(10) and component types per time unit constraints (11) and (14). Constraints (12) are in magnitude of time units. The size of constraints (13) is constant, i.e. number of resources \( |R| = 7 \). Hence the number of constraints is in magnitude \( O(|O| + |T||(|P| + 1)) \).

5.2. Problem complexity

Our production planning problem is NP-hard and we show this by reducing a special case of our problem to Knapsack problem. By Complexity Theory, reducing a special case of a problem to an NP-hard problem is sufficient to show the NP-hardness of that problem.

**Knapsack problem**

**Given** a set \( I \) of items, such that each \( i \in I \) has a weight \( w_i \) and a value \( v_i \), along with a maximum weight capacity \( W \). Find the subset \( I' \) of items that fits to the capacity, i.e. \( \sum_{i \in I'} w_i \leq W \), and maximizes the total value, that is \( \sum_{i \in I'} v_i \).

**Theorem 1.** **AME production planning problem can be reduced to Knapsack problem, hence it is NP-hard.**

**Proof.** The special case of our planning problem is as follows. There are two product types \( P_{N1}, P_{N2} \) and they are produced at workcenters \( wc_1 \) and \( wc_2 \) in one-step production process. The order set is \( \{ o_0, o_1, \ldots, o_k \} \) such that the product type required for \( o_0 \) is only produced at workcenter \( wc_1 \), and all other orders are produced at \( wc_2 \). The minimum required time horizon of \( wc_1 \) is \( H_1 \) and there is no minimum time horizon requirement of \( wc_2 \), numerically specified as zero. Deadlines are \( d_i = d_o = W > H_1 \) and latest start times are equal to deadlines for all \( i = 1, 2, \ldots, n \). The customer order delay tolerance is zero, so orders are either included with their entire amounts or rejected.

To satisfy the minimum time horizon of workcenter \( wc_1 \) order \( o_0 \) must be selected. Hence time horizon should be equal to \( W (=d_o) \). The only objective part left to consider is related to rejections. Minimizing the weighted rejected orders is equivalent to maximizing the selected orders weighted with revenues such that the total workload of processing orders \( o_1, \ldots, o_k \) stays within time \( d_0 \).

The solution to above order rejection decisions corresponds to solving a knapsack problem with capacity \( d_0 (=W) \), item weights correspond to process times of orders and item values correspond to the revenues of the orders.

**On the complexity of real-life problem instances.** As seen in the reduction to knapsack problem in Theorem 1, the complexity mainly comes from rejection decisions. In the real-life problem instances of AME company the ratio of rejected orders is small which makes it more hopeful to obtain optimal results within reasonable times using state-of-the-art MILP solvers like CPLEX and Gurobi. In the next section, we will experiment the computational performance of our model by generating a set of instances.

6. Computational experiments

The proposed MILP model is implemented in Python 3.8 using the solver Gurobi 8.0. This section will evaluate the computational performance of the model in order to determine the frequency in which the production plans can be updated. Two main characteristics that influence the computational performance of a model are the quality (i.e. tightness) of the linear relaxations, as well as the size of the model (Ku and Beck, 2016). Here the tightness can be defined as the closeness of an optimal solution of an Integer Linear Programming model with the optimal solution of its relaxation. The goal of our sensitivity analysis in this section is to explore the effect of the MILP model size on the computational performance.
6.1. Computational performance

In order to obtain reliable information about the computational cost of our MILP formulation, we generated a set of representative problem instances. These instances carry similar properties as the ones that are extracted from the information system of AME company.

Generating representative problem instances for the sensitivity analysis is done in two steps. First, we calculated statistical properties, means and standard deviations, of several parameters like the number of orders, the order quantities, and delivery deadlines of real-life problem instances (listed in Table 2). We then define either normal probability distributions with the same mean and variation values or the empirical probability distributions. For confidentiality reasons the precise values are not disclosed. Second the problem instances are generated by sampling these instance parameters from the defined distributions. An increment of 100 is selected for instance size, and an order range of [0, 600] is selected to find the computational time of the model, as seen in Fig. 6. The following section and later sensitivity analysis section report the results of real-life instances.

Two important properties of problem instances are varied in our sensitivity analysis; the size of order set \(|O|\), and the customer tolerance value \(\tau\). The former directly determines the number of constraints \((2)–(10)\) and the latter the number of \(\theta_{oi}\) decision variables. Other properties such as order composition, quantity, and delivery deadlines are generated through sampling from the aforementioned distributions.

The deadlines of the orders are sampled from the normal distributions for confirmed forecast orders and from empirical distributions for sales orders. Forecast orders are represent the predicted amounts of near future orders and these amounts are confirmed by the customers. They are specified in the ERP system monthly for each product type. The order quantities are sampled from the normal distributions that are defined by using historical orders per product type. The order quantity per order is scaled proportionally to the size of the order set \(|O|\) in the generated instance to preserve the ratio of the production capacity. The mean and standard deviation are found as in expression (17)

\[
\mu_{i,|O|} = \mu_{i} \frac{1}{|O|} \sum_{p \in \Pi} \frac{|O_{p,f}(i)|}{|H|}, \quad \text{where} \quad \mu_{i} = \frac{1}{|H|} \sum_{p \in \Pi} \frac{\sum_{o \in O_{p,f}(i)} q_{o}}{|O_{p,f}(i)|} \tag{17}
\]

where \(H\) is the set of historical problem instances and \(O_{p,f}(i)\) is the set of forecast orders for product type \(i\) in historical problem instance \(p \in H\). The standard deviations are obtained in the same manner as the mean.

Test 1: Varying the order set. This test investigates the effect of changing the size of order set \(|O|\), on the computational performance. The results are shown in Fig. 6(a). The red star on the graph shows the mean computational time of the instances with corresponding \(|O|\) values, while the blue bars show the standard deviation of the running time. As the amount of orders increased, the magnitude of variability of the solution time also increases.

Test 2: Varying the delay tolerance. This test investigates the effect of changing the value of customer order delay tolerance on the computational performance. Increasing the value of \(C\) causes an increase in the number of \(\theta\) binary integer decision variables of orders. Problem instances were generated for values of \(C\) between 0 and 4. The computational time required to solve these instances were plotted against the amount of orders in the instance, as shown in Fig. 6(b) in which the mean is shown by the red star while the blue bar shows the standard deviation.

The results show an exponential increase in running time and variability in running time as the value of \(C\) increases. This is expected, since the possible target delivery times of orders increase with higher values of delay tolerance parameter \(C\).

6.2. Real-life problem instances

We selected four AME real-life problem instances, each includes customer orders for the upcoming one year time period. The instances were extracted at the beginning of four different weeks 21, 27, 34, and 50 in year 2019. The selected weeks of the year are separate from each other due to avoiding the overlap in orders. The time limit of 9 hours is defined for the runnig time of the MILP model, as this can be the maximum running time for the daily basis use of the planning model. Table 2 shows that a near-optimal solution was found for each of real-life problem instances. The optimality gap value is an indication for the quality of the solution. An optimality gap near 0 indicates that the solution is nearly optimal, whilst an optimality gap near 1 indicates that no strong conclusion about the solution quality can be drawn. As seen in Table 2, only problem instance Week-50 has positive optimality gap with small value may cause slight deviation from optimality. This can be explained considering the length of planning horizon and the number of rejections of the instances. As seen in Table 3, Week-27 has planning horizon 70 days and 18 rejections, whereas, Week-50 has 90 days and 90 rejections (Fig. 8).

In order to see how optimality gap values stay in acceptable small values, i.e. providing near-optimal solutions, we solved our model for a range of time limits up to 900 min. The results of the 4 problem instances with varying time limits is shown in Fig. 7.

6.3. Sensitivity analysis: Planning Horizon

This section explores how the length of the planning horizon, i.e. \(\tau\) value, is influenced by different operational conditions. In our MILP model the system characteristics in capacity constraints (12) and minimum workload constraints (13) have strong impact on the main decisions regarding order selection, hence the planning horizon. Therefore we make the sensitivity analysis of the planning horizon by varying the minimum workload requirements and utilization ratios in the real-life problem instances.

Minimum workload requirements. The minimum workload requirement for AME problem instance Week-27 as introduced in the previous section was varied for five out of the seven resources considered. The other two resources (IM-2 and WIRE-0) are set at a minimum workload of 0 weeks (i.e. no minimum workload). The upper bound utilization ratio for each resource is 1 (i.e. each resource can be utilized for the amount of hours equal to the daily production capacity).

The results of the different combinations of minimum workload requirements are shown in Table 3. Initially the minimum workload requirement was set at 3 weeks for each of the 5 resources, as this is comparable to the current practice at AME. This is shown as Scenario 1 in Table 3. In scenarios 2–15 the minimum workload requirements of individual resources are reduced. As it is seen in Table 3, reducing the minimum workload requirements does not make any change in the selected \(\tau\) value and number of order rejections.

The selected planning horizon results in a specific selection of orders, each transferring workload to the required resource(s) to produce the order. The selected orders result in approximately 8.1, 10.9, 13.9, 7.8 and 5.5 weeks of workload for IM-0, IM-1, WIRE-1, SA-0 and PRD-12 respectively. Further increases in minimum workload requirements after scenario 16 change the length of the planning horizon. Scenarios 17–24 show the effect of increasing the minimum workload requirements by one or two weeks for each resource. Accordingly the value of planning horizon \(\tau\) as well as the amount of rejections change. It can be seen that \(\tau\) will only change when the minimum workload requirements exceed the amount of workload in the reference scenario.

The same performance tests were also conducted on the AME problem instance Week-50. This problem instance was chosen as it has a much higher amount of rejection; indicating that the system in general is overloaded and there are many more orders than the system can...
Table 2
Computational cost of solving the real-life instances within time limit of 9 h.

| Problem instance | \( |O| \) | \(|T| \) (days) | \( C \) (weeks) | Running time (minutes) | Optimality gap \( (\times 10^6) \) | Constraints \( (\times 10^7) \) | CVs\(^a\) \( (\times 10^5) \) | BIVs\(^a\) \( (\times 10^7) \) |
|------------------|--------|-------------|---------|---------------------|-------------------|-----------------|-----------------|-----------------|
| Week-21          | 1170   | 203         | 2       | 55.3                | 0                 | 4.76            | 2.32            | 2.66            |
| Week-27          | 1497   | 214         | 2       | 100.8               | 0                 | 5.05            | 2.45            | 2.90            |
| Week-34          | 1614   | 168         | 2       | 101.3               | 0                 | 4.02            | 1.92            | 2.41            |
| Week-50          | 1606   | 176         | 2       | 540.0               | 0.0057            | 4.20            | 2.02            | 2.52            |

\(^a\)CVs: Continuous variables, BIVs: Binary variables.

Table 3
Planning horizon sensitivity Test 1\(^a\).

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Input: Minimum workload requirement (weeks)</th>
<th>Output</th>
<th>( \tau )</th>
<th>Rejects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IM-0 3, IM-1 3, WIRE-1 3, SA-0 3, PRD-12 3</td>
<td>70</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>2,3,4</td>
<td>IM-0 2,1,0 3, IM-1 3, WIRE-1 3, SA-0 3</td>
<td>70</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>5,6,7</td>
<td>IM-0 3, IM-1 2,1,0 3, WIRE-1 3, SA-0 3</td>
<td>70</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>8,9,10</td>
<td>IM-0 3, IM-1 3, WIRE-1 2,1,0 3, SA-0 3</td>
<td>70</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>11,12,13</td>
<td>IM-0 3, IM-1 3, WIRE-1 2,1,0 3, SA-0 3</td>
<td>70</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>14,15,16</td>
<td>IM-0 3, IM-1 3, WIRE-1 2,1,0 3, SA-0 3</td>
<td>70</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>IM-0 8, IM-1 10, WIRE-1 13, SA-0 7, PRD-12 5</td>
<td>70</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>IM-0 9, IM-1 11, WIRE-1 14, SA-0 8, PRD-12 6</td>
<td>75</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>IM-0 10, IM-1 11, WIRE-1 14, SA-0 8, PRD-12 6</td>
<td>83</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>IM-0 9, IM-1 12, WIRE-1 14, SA-0 8, PRD-12 6</td>
<td>75</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>IM-0 9, IM-1 11, WIRE-1 15, SA-0 8, PRD-12 6</td>
<td>75</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>IM-0 9, IM-1 11, WIRE-1 14, SA-0 9, PRD-12 6</td>
<td>80</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>IM-0 9, IM-1 11, WIRE-1 14, SA-0 8, PRD-12 7</td>
<td>130</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>IM-0 10, IM-1 12, WIRE-1 15, SA-0 9, PRD-12 7</td>
<td>130</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>IM-0 11, IM-1 13, WIRE-1 16, SA-0 10, PRD-12 8</td>
<td>130</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Instance Week-27 with time length \( T_{\text{max}} = 150 \), delay tolerance \( C = 2 \).

handle. A time limit of 2 h was imposed. The optimality gap for each scenario was significantly low such that each solution can be considered as a (near-)optimal solution. The results can be seen in Fig. 8.

The time order selection in Scenario 1 with \( \tau = 90 \) resulted in a workload of 11.3, 15.4, 18.92, 16.9 and 11.6 for IM-0, IM-1,WIRE-1, SA-0 and PRD-12 respectively. In Scenario 2 in Table 3, the minimum workload is 11, 15, 18, 16 and 11 for IM-0, IM-1,WIRE-1, SA-0 and PRD-12 respectively; i.e. the amount of workload resulting from the order selection in scenario 1 rounded down to the nearest week. As in the problem instance Week-27, increasing the minimum workload to a value below the workload initially derived from the order selection does not affect the output of the model. Likewise, increasing any of the minimum workload requirements to more than the workload derived from the orders in the reference scenario results in an increase of the time horizon \( \tau \).

Utilization ratio. The utilization ratio is also an important factor for determining the planning horizon, \( \tau \) value, since it directly effects the production capacity. Reducing the utilization ratio of a resource implies reduction of the daily capacity of the resource (i.e. amount of hours that the resource can be filled with each day). The effect of the utilization ratio is investigated by varying values of it for the AME...
Fig. 8. Planning horizon sensitivity Test 2*
*Instance Week-50 with time length $T_{\text{max}} = 150$, delay tolerance $C = 2$.

Table 4
Planning horizon sensitivity Test 3*

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Input: Utilization ratio</th>
<th>Output</th>
<th>Rejects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IM-0 IM-1 WIRE-1 SA-0 PRD-12</td>
<td>$r$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1 1 1 1 1</td>
<td>70</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>1 0.5 1 1</td>
<td>70</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>1 1 0.5 1</td>
<td>70</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>1 1 1 0.5</td>
<td>70</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>1 1 1 0.5</td>
<td>70</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>0.5 0.5 0.5 0.5</td>
<td>70</td>
<td>49</td>
</tr>
<tr>
<td>7</td>
<td>0.3 0.3 0.3 0.3</td>
<td>70</td>
<td>111</td>
</tr>
<tr>
<td>8</td>
<td>0.2 0.2 0.2 0.2</td>
<td>70</td>
<td>203</td>
</tr>
</tbody>
</table>

*Instance Week-27 with time length $T_{\text{max}} = 150$, delay tolerance $C = 2$.

problem instance Week-27 with a minimum workload requirement of the 5 critical resources of 3 weeks. Table 4 shows several scenarios with varying upper bound utilization value combinations for the resources. It can be seen that different upper bound utilization values does not affect the value of the time horizon $r$, although it significantly influences the value of the amount of rejected orders.

6.4. Evaluation: Proposed approach vs. Current practice

This section aims to present the practical implications of the proposed production planning model. This is achieved through comparing the workload per resource that is derived from both order selection approaches; the approach presented in this paper with a dynamically selected time horizon and the current practice at AME with a fixed time horizon of one month (31 days). The resulting workload per resource is depicted in Figs. 9 and 10. The dark green color shows the minimum workload per resource, while the light green shows the amount of workload that is derived from the orders selected based on the given time horizon. The portions of the bar shaded in orange indicate a shortage of workload: this means that the given time horizon does not ensure for a selection of orders that is able to fill required minimum workloads.

It can be seen in Figs. 9 and 10 that the traditional approach does not ensure for the required minimum workload to be filled. Resource IM-2 for example only has sufficient workload for less than 1.5 weeks. This means that if a production plan is created at $t = 0$, this resource can make a planning for less than 1.5 weeks in advance. This is most likely quite undesirable, as generally employees need their working schedule for more than 1.5 weeks in advance. The same applies for the other resources with a workload shortage; the current method cannot ensure that in the upcoming scheduling period the workload is filled entirely. This means that these resources may not be utilized as efficiently as possible. The flexible time horizon approach shown in Fig. 10 on the other hand ensures that at least the desired minimum workload is filled. This means that a schedule for production that ensures optimal capacity usage can be made for at least the time period of this minimum workload per resource.

Fig. 9. Traditional approach with $r = 31$ fixed for planning instance Week-27. (Note: Different scale than Fig. 10).

Fig. 10. New approach with $r = 70$ decided by the workload based on minimum workload. (Planning instance Week-27).

7. Conclusions and future research

In this paper, a production planning model is proposed for a make-to-order type electronics manufacturing company. The proposed model adopts a workload-dependent time horizon approach for order selection. More specifically, the model converts customer orders into production targets, while accounting for the production system capacity as well as the desired workload amount. The model integrates order selection with production planning and capacity management by dynamically determining a planning time horizon (i.e., a workload-dependent order selection cut-off point) that flexibly adapts to the workload of the given order set. The use of the developed model is illustrated with a case study obtained from AME. Numerical experiments based on industry-scale problems show that the proposed model (i.e., planning with a workload-dependent time horizon) outperforms the current practice (i.e., planning with a fixed-time horizon). More specifically, the industry case study shows that the proposed model achieves the desired workload levels, although the current practice with fixed-time horizon may not necessarily achieve this target.

The proposed model is generic and applicable to make-to-order manufacturing companies with a high-mix nature. The model ensures that the desired resource capacity levels are achieved regardless of the
product mix in the order set. Computational experiments show that industry-size problem instances can be solved with a reasonable quality in two hours, making the model suitable for production planning activities on a daily basis. Consequently, the industry partner has decided to deploy the proposed approach in their Information Systems, and this project has been started recently via signed agreements.

We propose several directions for future research. The developed model can be used as a capacity management tool using aggregated product groups and predicted demand. For this purpose, machine-learning models could be trained to make high-quality demand predictions in a selected future time horizon. The model can be also extended to support tactical decisions regarding whether to expand (or reduce) capacity. Another direction for future research could develop heuristic approaches to solve the MILP model in shorter times. Decomposition techniques like Column Generation may be employed when a specific time limit and/or solution quality is desired. In addition, it may be interesting to expand the scope of the MILP model by including process quality changes, hence maintenance activities in the plan besides setup times. Another interesting direction may be to include and model human involved operations.

In hierarchical production control setting, our approach may be the first phase to be followed by a scheduling step and the second phase is currently being studied as a follow-up work of this paper.

CRediT authorship contribution statement

Murat Firat: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Supervision, Funding acquisition. Julie De Meyere: Data curation, Writing - original draft, Visualization, Software. Tugce Martagan: Conceptualization, Supervision, Writing - original draft, Writing - review & editing, Funding acquisition. Laura Genga: Conceptualization, Writing - review & editing.

Acknowledgments

This research is partially funded by TUE-AME collaboration agreement. We are thankful to Dirk van Driel from Applied Micro Electronics “AME” B.V. for sharing the information of their production system, providing us their business insights as well as the data set used for our analysis.

References