Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Dynamic Aggregate Planning of a Make-to-Order Production System

Valtteri Vironmäki
Joonas Ahola
Esko Niemi
esko.niemi@aalto.fi

Aalto University School of Engineering

Research Article

Keywords: Dynamic aggregate production planning, make-to-order production, mathematical optimization, resource constrained project scheduling problem, rolling horizon

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11261258/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Abstract

In Make-to-Order (MTO) production, products and production are non-standard and incoming orders arrive variably, with varying work content and due dates. Aggregate Production Planning (APP) takes place in the medium run, and the purpose of it is to match resource capacity with demand. In APP, orders are roughly scheduled for processing so that the resource capacity constraints and deadlines can be met. The static resource constrained project scheduling problem is the conventional research approach, but real-world production planning requires dynamic scheduling to integrate new orders over time. The dynamic model in this study employs periodic scheduling, treating it as sequential static mixed integer linear optimization problems, focusing on minimizing tardiness and rescheduling disruptions. The model is applied in experiments concerning a Flexible Manufacturing System (FMS) -centered MTO production system. The experimentation focuses on the effects of replanning frequency, order arrival rate, delivery date slack, forward planning vs. backward planning, restrictions in order processing activity timing and activity workload allocation method on the planning results. Based on the results, conclusions and guidance applicable to practical production planning are provided.

1 Introduction

Production planning is vital for maintaining a competitive edge, ensuring the availability of essential resources such as raw materials, workforce, and machinery to produce goods according to schedule. In Make-to-Order (MTO) production, Aggregate Production Planning (APP) is often handled as a project scheduling challenge, using models like the Resource-Constrained Project Scheduling Problem (RCPSP) or its multi-project variant, the Resource-Constrained Multi-Project Scheduling Problem (RCMPSP). However, static RCPSP models fall short in real-world settings where new orders arrive continuously and need integration into existing plans. This dynamic aspect calls for a dynamic aggregate production planning approach, as unpredictable arrival times and diverse order content affect system efficiency directly. Efficient scheduling of new orders is critical, but optimal timing – whether at regular intervals or triggered by a specific arrival threshold – remains challenging. Furthermore, while rescheduling enhances resource utilization, it can disrupt production stability. Typically, orders can only be delayed because advancing them often requires also advancing customer order engineering or raw material procurement, which are usually not feasible.

The focus in this study is primarily on minimizing order tardiness and rescheduling, i.e., optimizing system's operational efficiency rather than financial costs or Work-in-Progress (WIP). To improve dynamic production planning, this study seeks to answer several research questions:

- What is the optimal frequency for scheduling incoming orders – and rescheduling existing orders?
- How do average order arrival rates impact tardiness and the need for rescheduling?
- How does delivery date slack influence key performance metrics?
- What are the effects of forward and backward planning methods on system performance?
- How do constraints on order processing activity timing and activity workload allocation method affect planning outcomes?

To address the research questions, we apply a Mixed-Integer Linear Programming (MILP) model to conduct factorial experiments across various parameter sets. The model incorporates a rolling horizon planning approach, periodically updating the production schedule at predetermined intervals to integrate new orders. Consequently, the framework processes new order scheduling and potential revisions to existing orders as a sequential series of static planning problems.

1.1 Literature Review

Aggregate production planning is often treated as a project scheduling problem, formulated as a Resource-Constrained Project Scheduling Problem (RCPSP). The RCPSP is widely applied in contexts like construction and MTO manufacturing systems, where the goal is to optimize project duration amidst constraints like resource availability [1, 2]. The version of the problem in which multiple projects are scheduled simultaneously and compete for the same limited resources, consist of various activities, and may have different start times and due dates, is referred to as the Resource-Constrained Multi-Project Scheduling Problem (RCMPSP) [3]. Traditional practical planning systems, such as Manufacturing Resource Planning (MRP), create aggregate plans with weekly or monthly resolutions but rely on standardized schedules. Project scheduling-based approaches can offer flexibility by scheduling activities individually per order to achieve shorter lead times with a focus on achieving leveled resource profiles through workload allocation [4]. The intensity of activities can either be fixed or variable across time periods. Fixed intensity means that the activity's workload is evenly distributed over its planned duration. Variable intensity, on the other hand, allows different amounts of workload to be allocated to the activity across different execution periods. Variable intensity enables a more balanced resource profile, i.e. higher utilization of capacity. In order for variable intensity to guarantee feasibility, the resource types used are assumed to be continuously divisible – an assumption that typically holds for a group of workers but may not always apply to a single machine or a machine group [5].

In production environments, dynamic scheduling often occurs due to unpredictable events [6, 7, 8]. A rolling horizon approach iteratively solves decision-making problems by updating plans as new data becomes available, allowing adjustments to real-time events [9]. Fang and Xi [10] applied this to job-shop scheduling problems. Dynamic environment referred to a situation where new jobs arrived randomly over time, machines could break down, and job due dates could change during processing. To address this, the authors proposed both a periodic and an event-driven rolling horizon scheduling method. The job scheduling was performed using a hybrid algorithm combining a genetic algorithm with a dispatch rule. The objective was to minimize the makespan as well as job tardiness. The rolling horizon approach enables the solution of real-time scheduling problems while also reducing the computational complexity, as only a limited subset of the overall problem is optimized at a time. The rolling horizon scheduling strategy implemented in this study is based on the approach presented by Fang and Xi.

Several studies have examined dynamic RCMPSP scheduling approaches based on heuristic rules using simulation, e.g. [11] and [12]. Later on, optimization studies have appeared: [13], [14], [15], [16] and [17]. However, their focuses have been on developing optimization algorithms for specific problems, and the testing has been limited to selected case problems. The papers do not offer much guidance to industrial production managers in the way aggregate planning in MTO environments should be executed. This paper aims to address this gap by testing the effects of decision parameters that the managers can influence under the circumstances assumed in this study. The ultimate goal is to better understand how various replanning and rescheduling strategies in dynamic MTO environments behave and provide practical guidelines for industrial decision makers. To our knowledge, the important and practical research problems presented above have not been studied before in a systematic manner like we do.

2 Model

The model in this study is based on the resource-constrained multi-project scheduling problem, aiming to efficiently allocate multiple orders to resources with fixed daily capacities. Orders are highly variable, and each order comprises sequential activities, which must be completed according to precedence constraints. Workload, measured in hours, dictates the activity duration, which includes additional slack beyond the workload-derived time. The dynamic nature of the problem stems from the continuous but fluctuating arrival of new orders, necessitating regular plan updates every day or every few days. The planning starts with an empty factory, continuing until the end of the planning horizon when all orders are processed, featuring initial and final phases where capacity is actually not a constraint. Each planning iteration creates constraints for the next schedule. The rescheduling approach combines reactive scheduling with robust features

to optimize production efficiency (minimizing tardiness) while reducing deviations from the previous plan. Due dates depend on activity durations and slack coefficients. Future orders are unpredictable, preventing pre-scheduling.

The mathematical formulation of our APP optimization problem is similar to e.g. the static model of Morton and Pentico [18]. There are some differences due to the way the rolling horizon method is implemented in this model, which affects certain constraints. Table 1 lists the symbols used in our mathematical model. Among these $c_{i,t}$, the activity completion times, are the decision variables that govern the solution. In this context, an "activity" refers to a single activity within an order. The set *I* thus includes all individual activities in all orders, and each activity is logically linked to its corresponding order through precedence constraints.

Table 1
Model symbols

Variable	Description
C_i	Finish time of activity i
$c_{i,t}$	1, if activity i is finished at time step t, 0 otherwise
$X_{i,t}$	1 if activity i is ongoing at time step t, 0 otherwise
$R_{i,k,t}$	Workload of activity i on resource k at time step t
$tardiness_i$	Tardiness of activity i
maxtardiness	Maximum tardiness
$scheduled_i$	Planned starting time of activity *i*
$rescheduled_i$	1, if activity i is rescheduled at current iteration, 0 otherwise
fw_i	Forward planning: starting time of activity i from its arrival
bw_i	Backward planning: finish time of order's last activity from its due date
Parameter	
P_i, A_i, D_i	Duration, arrival day, and due date of activity i
I	Set of activities
T	Final time
K	Set of resources
$capacity_k$	Maximum capacity of resource k
$WC_{i,k}$	Work content of activity i on resource k
timestep	Current time step
M	Big *M*

Mathematical Formulation. The constraints (1 − 18) are presented first and the objective (19) later.

$$\sum_{t=1}^{T} c_{i,t} = 1, \ \forall\, i \in \ I$$

1

$$\sum_{t=1}^{T} tc_{i,t} = C_i, \forall\, i \in I$$

2

$$C_i - P_i \geq A_i, \forall\, i \in I$$

3

$$X_{i,t} = \sum_{u=t+1}^{t+P_i} c_{i,u}, u \leq T, \forall\, i \in I, \forall\, t \in 1 \ldots T$$

4

$$R_{i,k,t} = \frac{WC_{i,k} * X_{i,t}}{P_i}, \forall\, i \in I, \forall\, k \in K, \forall\, t \in 1 \ldots T$$

5

$$\sum_{t=1}^{T} R_{i,k,t} = WC_{i,k}, \forall\, i \in I, \forall\, k \in K$$

6

$$R_{i,k,t} \leq M * X_{i,t}, \forall\, i \in I, \forall\, k \in K, \forall\, t \in 1 \ldots T$$

7

$$C_j = C_i - P_i, \forall\, j \in P(i), \forall\, i \in I$$

8

$$C_j \leq C_i - P_i, \forall\, j \in P(i), \forall\, i \in I$$

9

$$capacity_k - \sum_{i \in I_o} R_{i,k,t} \geq 0, \forall\, k \in K, \forall\, t \in 1 \ldots T$$

10

$$C_i - P_i \geq timestep, \forall\, i \in I$$

11

$$C_i - D_i \leq tardiness_i, \forall\, i \in I$$

12

$$maxtardiness \geq tardiness_i, \forall\, i \in I$$

13

$$C_i - C_{i'} \leq M * rescheduled_i, \forall\, i \in I_s$$

14

$$C_i - scheduled_i - P_i \geq 0, \forall i \in I_s$$

15

$$fw_i \geq C_i - P_i - A_i, \forall i \in I$$

16

$$bw_i \geq D_i - C_i, \forall i \in I_l$$

17

$$bw_i \geq C_i - D_i, \forall i \in I_l$$

18

$$Min \quad w_1 maxtardiness + w_2 \sum_{i \in I} tardiness_i + w_3 \sum_{i \in I} rescheduled_i + w_4 \sum_{i \in I \text{ or } i \in I_{last}} (fw_i \text{ or } bw_i)$$

19

Constraint (1) states that an activity must finish once and only once. Finish time of an activity is set by constraint (2). Activities cannot start before arrival time of corresponding order (3). Activity i is ongoing at time step t, if variable $X_{i,t}$ = 1 (4). The load caused by an activity on a resource is distributed evenly over the duration of the activity by constraint (5). Constraints (6) and (7) allow the workload to be distributed freely across the days during which the activity is ongoing. Either constraints (5) or (6) and (7) are used depending on which type of workload allocation is modeled.

The timing template and precedence constraints are closely related. There are two types of timing templates: in the free timing template, there may be a waiting time between execution of consecutive activities of an order, whereas in the fixed timing template, a gap is not allowed. Precedence constraints (8) ensure that the next activity of an order starts immediately after the previous one has finished. Similarly, the constraints (9) must be imposed when using the free timing template. Here $j \in P(i)$ denotes the set of immediate predecessors of activity i. The resource capacity must not be exceeded at any time (10). I_o is the set of activities that consists of both activities scheduled in the current planning iteration and activities that are still ongoing on the resource from the previous iteration. The start of an activity cannot be set in the past (11). Tardiness and maximum tardiness are set in (12) and (13). Rescheduling is tracked using constraint (14), where $C_{i'}$ is the completion time of activity i from the previous iteration. Moving activities backwards in time is prevented with constraint (15), where I_s is set of scheduled activities, and $scheduled_i$ is planned starting time of activity i from previous scheduling iteration. To accomplish forward planning and backward planning, variables fw_i are used to express the difference between the activity's starting time and its arrival time (16) and variables bw_i to express the absolute value of the difference between the order's last activity's finish time and its due date (17 and 18). Either one of these variables is contained as a penalty in the objective (19) depending on which planning method is used, along with the primarily minimized weighted maximum tardiness, total tardiness and number of reschedulings.

3 Experiments

To answer the research questions, we use our model to test the effects of the test parameter values on a hypothetical manufacturing system with pre-generated random order data. The data is incrementally "revealed" to the model over time mimicking the flow of incoming orders to a real manufacturing system. The dynamic planning problem is thus solved as a series of static problems. Every time a scheduling occurs, a new model instance is created and solved. The schedule of each iteration (planned start times of the activities) is saved and compared in the next replanning iteration. Each model considers only variables that belong to the activities that have arrived but have not yet started. This implies that

information about earlier models must be taken into account when creating a new model to respect precedence and capacity constraints correctly.

In the following sections, the word replanning (or planning) refers to the act of updating the production schedule, whereas the word rescheduling refers to the change in start time of a previously scheduled activity caused by replanning.

The model is implemented in Julia (version 1.11.0), and the JuMP modeling language [19] is used for optimization, together with the HiGHS solver [20].

3.1 Manufacturing System

The schematic of the manufacturing system is shown in Fig. 1. The system consists of one Preprocessing center (Pre), one Flexible Manufacturing System (FMS), two Postprocessing centers (Post1 and Post2), one Quality Control center (QC), and finally a Dispatch center (Dispatch). This type of FMS-centered manufacturing system is common in parts manufacturing in the mechanical engineering industry.

The process path of an order (i.e., a product to be manufactured) follows one of twelve alternative paths. The probability of a specific path is calculated based on the proportions shown in Fig. 1. It is worth noting that only the FMS and the dispatch center are included in the process path of every single order.

3.2 Data Generation and Design of Experiments

Although the model itself is deterministic, the generated data sets incorporate stochastic elements. This randomness is intended to reflect the variability of a real-world production environment, where both the structure of incoming orders and the duration of individual processing activities can vary.

The interarrival time of the orders is exponentially distributed. The resource that an activity loads is determined by the order path. The work content hours of the activities are modeled as realizations from a truncated gamma distribution. The distribution parameters depend on the specific resource that performs the activity and are presented in Table 2. The rate and scale parameters of the gamma distribution can be derived from the mean and the Coefficient of Variation (CV). The allowed duration of the time within which the activity must be executed depends on the duration coefficient and the shift arrangement of the resource according to Eq. (20), except resource Dispatch, where the duration is always 2 days.

Table 2
Activity work content and duration parameters

Parameter	Pre	FMS	Post1	QC	Post2	Dispatch
Mean work hours	5	5	5	8	10	1
Work hours CV	0.5	0.3	0.5	0.5	0.5	0.5
Duration coefficient	3	5	3	3	3	N/A
Shift arrangement, hours	8	24	8	8	8	N/A

$$Duration = \lceil Duration\ coefficient * Work\ content / Shift\ arrangement \rceil \quad (20)$$

The average nominal flow time of the orders (with the fixed timing template) is 8 days, calculated from the values provided in Fig. 1 and Table 2.

The due date of an order is determined by the duration of its activities and the slack coefficient. More formally:

$$Due\ date = \lceil (1 + Slack\ coefficient) * \Sigma Duration \rceil + Order\ arrival\ date \quad (21)$$

The time horizon of the model, i.e. planning horizon, is 100 days (and some cooling period after that), and thus the number of orders is determined by the arrival rate of orders. The arrival rate, one of the model parameters, here means the average number of orders arriving in one day.

Capacity is set on the basis of the average order arrival rate, the average workload of activity on certain resource, and the desired utilization rate:

$capacity_k = \lceil Order\ Arrival\ Rate * AvgWC_k / Utilization \rceil$ (22)

The utilization rate of 0.8 is used. This value is determined experimentally; it ensures that the problem is not trivial (i.e., no tardiness at all), while also keeping the computational complexity at a manageable level.

We implement a factorial experiment with the factors and their levels as follows:

- Planning frequency: 1 day, 2 days, 3 days
- Average order arrival rate: 3 orders/day, 5 orders/day
- Due date slack coefficient: 0.5, 0.75, 1.00
- Planning method: Forward, Backward
- Timing template fixed and free
- Free workload allocation within activity's duration with the free timing template

Therefore, this is a 3 x 2 x 3 x 2 x 3 experiment with a total of 108 combinations. A single trial consists of ten repetitions for each combination of parameter values and a single planning horizon. Thus, the total number of experiments is 1080 and the total number of orders in each trial is 300 or 500, depending on the average order arrival rate parameter. The planning takes place on average every other day, and therefore the optimization problem is solved approximately: 50 x 1080 = 54 000 times. The solver tolerance is left at 20%.

The linear scalarization method is used in the multi-criterion optimization problem. The objective function weights used are $w_1 = 10$ (maximum tardiness), $w_2 = 1$ (total tardiness), $w_3 = 1$ (number of reschedulings), and $w_4 = 0.01$ (forward planning) or $w_4 = 0.035$ (backward planning). The weights were determined empirically, with the aim of finding pareto efficient values without disproportionate emphasis on any individual objective.

4 Results and Discussion

The combinations of parameter values for planning frequency, order arrival rate, due date slack coefficient and planning method used in the experiments are presented in Appendix 1. These combinations are tested with fixed and free timing templates and with variable workload distribution together with the free timing template. The analyzed and recorded dependent values are the maximum tardiness, total tardiness and number of reschedulings that are minimized in the objective function of the planning problem. In addition total rescheduling length of all activities is reported and the individual activities rescheduling length analysed. Average totals of these dependent values from all orders in each repetition of each parameter value combination and their standard deviations (of the 10 experiment repetitions) are presented in Appendices 2–5. The main results are presented in Sections 4.1–4.3. Further analysis is provided in Section 4.4.

4.1 Maximum Tardiness

Fig. 2 shows the maximum tardiness for all trials containing each individual independent variable separately as a boxplot. There are 360 – 720 values in each boxplot depending on the variable combination.

It should be noted that in each case, all other independent variables are not fixed, that is, their values vary. Therefore, the variation shown is very large. However, the figure indicates that there are some clear relationships between the independent variables and the maximum tardiness:

- Forward planning results in lower maximum tardiness values than backward planning
- Relaxing the fixed activity timing reduces maximum tardiness and allowing variable workload distribution within activity duration further increases this effect
- Planning more frequently reduces maximum tardiness
- A higher order arrival rate reduces maximum tardiness
- A higher due date slack reduces maximum tardiness, but at a decreasing rate.

Multiple linear regression analysis is performed on the experimental results to determine which factor levels are significant with respect to the dependent variable. Replanning frequency, order arrival rate and due date slack are taken to be quantitative variables and planning method timing type and workload distribution are categorical variables in the model. Free workload distribution only appears combined with the free timing. Its effect should be interpreted to represent the effect additional to the effect of free timing. The linear regression is fitted to standardized data, so that the regression coefficients can be compared, see Table 3. Only the standard error is calculated from the non-standardized data, so that its value can be evaluated in the units of the dependent variable used e.g., in Figs. 2 – 4.

Table 3
Regression analysis results. Standard errors are given per order so that they can be compared with results in Figs. 2–4

Regression coefficients / P-values for standardized data	**Dependent variable**			
Independent variable	Maximum tardiness	Total tardiness	Reschedulings	Rescheduling move length
Planning frequency	0.15 / 0.000	0.06 / 0.018	-0.08 / 0.001	-0.01 / 0.663
Order arrival rate	-0.27 / 0.000	-0.26 / 0.000	-0.35 / 0.000	-0.33 / 0.000
Due date slack	-0.19 / 0.000	-0.15 / 0.000	-0.09 / 0.000	-0.06 / 0.011
Forward vs. backward planning	-0.33 / 0.000	-0.26 / 0.000	-0.23 / 0.000	-0.07 / 0.006
Free vs. fixed timing	-0.56 / 0.000	-0.32 / 0.000	-0.25 / 0.000	0.26 / 0.000
Free workload vs. even	-0.11 / 0.000	-0.07 / 0.017	-0.18 / 0.000	-0.52 / 0.000
R^2	0.63	0.29	0.33	0.32
Standard Error (non-standardized)	0.96 days	0.19 days	0.07	0.74 days

The results show that all independent variables are statistically significant in relation to maximum tardiness. Allowing activities to be scheduled freely – respecting precedence constraints and only forward in time – has the largest positive effect on maximum tardiness. Free workload allocation improves the results further. However, in the experimentation, this is only tested with free timing. Forward planning instead of backward planning has the next largest maximum tardiness reducing effect. Larger inflow of orders and looser schedules, i.e., larger due date slack, are less contributing factors and planning more frequently helps marginally.

To examine the interactions between the independent variables, a 2-way ANOVA is executed. Among the interaction effects, the interactions between the free timing (with and without free workload distribution) and order arrival rate, planning method and slack are statistically significant. Based on the means plots, it can be interpreted that free timing together with higher order arrival rate will decrease maximum tardiness less than the combined effect of free timing and higher order arrival rate values individually. Another interpretation is that free timing is more important with low order arrival rates, backward planning, low due date slack and infrequent planning than with high order arrival rates, forward planning, high due date slack and frequent planning. Similarly forward planning reduces maximum tardiness significantly more efficiently with low order arrival rates than with high rates compared to backward planning.

4.2 Total tardiness

For total tardiness analysis, as well as for number of reschedulings and total rescheduling length, the results are normalized by dividing them by the particular number of orders (300 or 500) to make them comparable. Figure 3 shows the total tardiness in average days per order as a boxplot. From the plots can be concluded that the variation in the values is large and the distributions are more skewed, i.e. the datapoints are concentrated closer to zero than for maximum tardiness. The regression R^2 (Table 3) value is low, and the frequency of replanning and free workload distribution are no longer statistically quite as significant as they were in the case of maximum tardiness. Generally, the regression main effects are in line with those to maximum tardiness and the observations above concerning the maximum tardiness boxplot are also similar.

According to 2-way ANOVA, significant interactions again exist between timing and other factors so that with free timing (and variable workload) the tardiness increasing effect of lower values for low order arrival rates, backward planning, low due date slack and infrequent planning tend to be largely absorbed. The same effect exists between forward planning and low order arrival rates. Also, the effect of low order arrival rate on total tardiness is efficiently eased by higher due date slack.

4.3 Reschedulings

Figure 4 shows the activity reschedulings per order. Contrary to the effect on tardiness, less frequent replanning seems to reduce the number of reschedulings, while the other effects have the same direction as for tardinesses. This is confirmed by the regression analysis. In Table 3, the coefficient sign for planning frequency is negative. Again, all independent variables are observed to be statistically significant. Higher average order arrival rate per day and free timing reduce the number of reschedulings mostly. Based on 2-way ANOVA results, significant interactions exist between timing and order arrival rate, planning frequency and planning method and between planning method and order arrival rate. Again, free timing has an enhanced effect on reducing reschedulings when order arrival rate is low or backward planning is used and in this case with high planning frequency. Also forward planning has a significant improvement effect on low order arrival rates.

Total rescheduling length has a high correlation with the number of reschedulings. The average value of individual activity move is 1.12 days with a standard deviation of 0.9 days. From regression analysis figures in Table 3 can be seen that planning frequency does not have significant effect on the length of the reschedulings. Free workload distribution together with free timing correlates with short moves. On the other hand, free timing alone produces longer moves. High order arrival rate significantly reduces length of reschedulings whereas forward planning and high due date slack have only a minor reducing effect.

4.4 Further Discussion

Reliability

The results presented reveal clear relationships between almost all independent variables and the various dependent variables. This is supported both by the visualizations derived from the data and by the statistical significance of the regression models. However, there is considerable variation in the results, as seen in the box plot figures and in the unexplained variance of the regression models (see the R^2 values in the regression model summary in Table 3).

Some of this variation can be easily attributed to the randomness inherent in the datasets, which could be better controlled by using a within-subjects experimental design instead of a between-subjects design. In addition, the allowed solver tolerance was quite high, which undoubtedly affected the precision of the results.

The large variance and presence of outliers, especially in total tardiness and the number of reschedulings, may also be due to the nature of these dependent variables: both are cumulative results of multiple planning iterations, and thus any variation introduced by randomness in individual iterations compounds in the final outcome. This issue is less prominent with maximum tardiness, as can be seen both visually (Fig. 2) and from the standard deviation statistics (Appendices 2–5). Total tardiness is recorded as sum of tardinesses of all activities, not the sum of tardiness of all orders. Depending on the order path and tardiness of the order, this activity-based total tardiness overemphasizes the tardiness of some orders.

The credibility of the results is also influenced by the use of the linear scalarization method in the multi-objective optimization model. Although this approach is relatively straightforward to implement, identifying suitable and Pareto efficient weights for objectives of different magnitudes is not a trivial task. Although some analysis was conducted with different weight combinations, the analysis was not exhaustive. An alternative multi-objective optimization method might have yielded more reliable results.

Results

Planning method. Based on the results, forward planning is the better option to minimize tardiness. The likely reason for this is the combination of the implementation of backward planning and the forward scheduling strategy. When orders are scheduled to finish exactly or close to their due dates, and they can only be rescheduled by moving them forward, this inevitably leads to tardiness when new urgent orders arrive at unfavorable times, with resources being overloaded. In practice, however, advancing orders could sometimes be allowed, if the orders are initially scheduled far in the future. In addition, scheduling orders to finish just on time might not be the way in which the planning is implemented in practice. It must be underlined that in this model, the focus was on minimizing tardiness while aiming for a stable schedule. Therefore, the model does not take into account the other benefits and drawbacks of different planning methods.

With forward planning, rescheduling is avoided much more often. Unlike forward planning, with backward planning, orders are scheduled for a longer period of time, making it theoretically possible to reschedule them at all. It is also observed that when rescheduling is necessary, the shifts are on average not very long and that the shifts are generally about the same length for both planning methods. This is somewhat in conflict with how one might expect the phenomenon to behave: with backward planning, even a short shift can easily result in tardiness, but with forward planning, this is not necessarily the case. This might lead one to assume that the shifts with backward planning would be shorter than with forward planning. However, this seeming contradiction can be explained by the fact that with forward planning, the need for rescheduling simply does not arise as often, which is why their rescheduling shifts also remain relatively short.

Timing template and workload distribution. Use of timing method that allows free timing of activities while respecting precedence constraints reduces tardiness and also number of reschedulings efficiently. This type of increased flexibility makes resource utilization much easier in the highly variable circumstances typical to MTO production. Use of this flexibility results in increased rescheduling lengths, as indicated by the positive coefficient in Table 3. The possibility to

allocate work content freely over the allowed activity duration further improves resource utilization, even so that both the number and length of reschedulings are reduced; again, see regression coefficients in Table 3.

Planning frequency. The tardiness is observed to increase when replanning is performed less frequently. This is an obvious finding, as in a dynamic production environment, infrequent replanning may not respond effectively to new orders, potentially resulting in uneven resource utilization. In particular, infrequent replanning can be disadvantageous for urgent orders.

However, less frequent replanning reduces the number of reschedulings. This is a rather expected result, since in less frequent planning, orders remain fixed for a longer period of time, which means that their start times cannot be adjusted as often. The quality of planning may also improve due to the larger number of orders being considered simultaneously. With daily replanning, jobs are not really fixed at all before they begin. Especially if orders are scheduled close to their arrival time (i.e., under forward planning), they may not be rescheduled at all. With fixed timing template rescheduling of later activities after the first activity has been started does not take place. Since minimizing tardiness and the number of reschedulings are conflicting objectives when the replanning frequency is a parameter, the optimal replanning frequency depends on the decision-maker's preferences.

Order arrival rate. With a higher average daily order arrival rate, both tardiness and the number of reschedulings are lower. The reason for this is likely the more efficient use of resources, made possible by handling more orders simultaneously. As a result, the effect of randomness of the workload decreases, making it easier to respect the capacity constraints.

The effect of randomness is also observed in the distributions of the dependent variables for different order arrival rates. In Figs. 2 – 4, a much larger dispersion can be seen in the results when the arrival rate is three orders per day. This suggests that randomness in the workload has a significant effect on the planning outcome.

Due date slack coefficient. A higher slack coefficient reduces tardiness. This is an obvious result: with a large slack coefficient, the due date of orders is farther away, making scheduling within capacity constraints easier. However, increasing the slack coefficient has diminishing returns. This is because, after a sufficiently large slack coefficient value, many orders can be scheduled with almost no tardiness, so the benefits cannot grow further. However, tardiness does not go to zero due to urgent orders or due to factors such as an exceptionally large number of orders arriving in a short period of time, or exceptionally large workloads of activities.

4.5 Practical considerations

Some comments on the results obtained and analysis above considering common industrial practice are provided.

The advantages of free activity timing and workload allocation on activity durations are obvious. The reason for comparing these to the use of fixed timing templates is that according to the authors' experience and knowledge, the latter are indeed commonly used in the industry. This practice is probably partly due to traditions and legacy software (MRP/ERP - Manufacturing/Materials/Enterprise Resource Planning, MES - Manufacturing Execution System, etc.), but also because managing often chaotic MTO manufacturing systems asks for order and administrative efficiency brought in by product families and related standard manufacturing recipes including standard timing templates. On the other hand, modern APS (Advanced Production Planning System) software offer tools for implementing the kind of practices this study suggests. Controlling the potentially increasing complexity sets additional challenges for production management, but as shown, the rewards are significant.

Another generally beneficial and controllable factor in our analysis is the forward planning method. However, minimization of WIP and the related tied capital favors timing the manufacturing operations as late as possible, i.e. backward planning. Another practical reason for favoring backward planning is that it provides more room for changes in

customer orders. Make-to-order production typically takes place in production of non-standard products, where customer needs tend to change even after the products are ordered. Being able to react to such requests creates a lot of goodwill at the customer, to say the least.

An illustrative detail concerning limiting excessive WIP and fluctuation in process flow is setting limits on the time gaps between activities or total flow time of an order. These sorts of constraints are very easy to implement in the kind of optimization models we are suggesting.

Our model is deterministic, i.e. no disruptions due to machine failures, worker absenteeism, material shortages, quality problems, etc. take place and work contents are exactly known at the time of planning. Robustness against uncertainty related to such incidents is partly reflected in the activity duration to work content ratio – *Duration coefficient* – of 3 for 1-shift work and 5 for 3-shift work. Shorter-leadtime aggregate planning results could be achieved with lower ratios and the resulting disaggregated schedules would still be theoretically mostly feasible [5]. In practice the appropriate tightness of the schedules is largely dependent on the reliability of the process, and the values of the duration coefficients must be determined locally by the production management.

The due dates of the orders were generated mechanistically based on the estimated (deterministic) manufacturing lead time. In reality, setting the delivery time of the orders is a result of a negotiation process between the customer and the producer. However, from the results of the current study one can conclude that the conclusions should be generally valid for any due dates that are not shorter than those set in the experiments.

5 Conclusions and further study

Under the given experimentation environment and the related assumptions, based on the results and their analysis, the main conclusions are summarized as follows:

- Relaxing the fixed activity timing reduces tardiness and need for rescheduling
- Allowing variable workload distribution within activity duration further increases these positive effects
- Forward planning generally results in lower tardiness values and less rescheduling than backward planning
- Tardiness rises with less frequent replanning, while rescheduling is reduced
- Higher order arrival rate, with capacity maintained equally, reduces tardiness and the need for rescheduling.
- Higher due date slack reduces tardiness, but at a decreasing rate
- Generally, the positive effects of free activity timing and forward planning are so dominant, that the negative effects of infrequent replanning, low order arrival rate and small due date slack are relatively reduced compared to what they are with fixed activity timing templates and backward planning.

These ideas could and should be noted by practitioners and tested in real life. Due to the nature of the problem, the probability and consequences of any risks in such testing should be limited. A further study direction could be to experiment with different types of production. E.g., study of larger and more complicated production systems might reveal interesting new insights. This direction might require using efficient heuristic solving algorithms, which should not be a major problem.

Another interesting minor issue is the forward/backward planning, which could be finetuned to cover planning of known but not released orders and using safety margins in backward planning.

Declarations

Competing interests

The authors report there are no competing interests to declare.

Funding

This work was supported by the Business Finland under Grant TwinFlow.

Author contributions

All authors contributed to the study conception and design. Valtteri Vironmäki created the experimentation system. Design of experiments, experiment execution and analysis were performed by Valtteri Vironmäki and Joonas Ahola. The first draft of the manuscript was written by Esko Niemi and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

1. Alfieri A, Urgo M (2015) Project Scheduling for Aggregate Production Scheduling in Make-to-Order Environments. In: Schwindt C, Zimmermann J (ed) Handbook on Project Management and Scheduling Vol. 2. Springer Cham, pp 1249–1266. https://doi.org/10.1007/978-3-319-05915-0
2. Hartmann S, Briskorn D (2022) An updated survey of variants and extensions of the resource-constrained project scheduling problem. Eur J Oper Res 297:1–14. https://doi.org/10.1016/j.ejor.2021.05.004
3. Sánchez MG, Lalla-Ruiz E, Gil AF, Castro C, Voß S (2023) Resource-constrained multi-project scheduling problem: A survey. Eur J Oper Res 309:958–976. https://doi.org/10.1016/j.ejor.2022.09.033
4. Toivonen V, Ikkala K, Niemi E, Korhonen A (2006) Planning and Scheduling System for Make-To-Order Production. CIRP Ann 55:493–496. https://doi.org/10.1016/S0007-8506(07)60466-1
5. Niemi E, Toivonen V (2008) Transformation of levelled production plans into detailed schedules in make-to-order manufacturing. Int J Serv Oper Inf 3:15–27. https://doi.org/10.1504/IJSOI.2008.017702
6. Ouelhadj D, Petrovic S (2009) A survey of dynamic scheduling in manufacturing systems. J Sched 12:417–431. https://doi.org/10.1007/s10951-008-0090-8
7. Tokola H, Ahlroth L, Niemi E (2014) A comparison of rescheduling policies for online flow shops to minimize tardiness. Eng Optim 46:165–180. https://doi.org/10.1080/0305215X.2012.753438
8. Baykasoğlu A, Karaslan FS (2017) Solving comprehensive dynamic job shop scheduling problem by using a GRASP-based approach. Int J Prod Res 55:3308–3325. https://doi.org/10.1080/00207543.2017.1306134
9. Sahin F, Narayanan A, Robinson EP (2013) Rolling horizon planning in supply chains: review, implications and directions for future research. Int J Prod Res 51:5413–5436. https://doi.org/10.1080/00207543.2013.775523
10. Fang J, Xi Y (1997) A rolling horizon job shop rescheduling strategy in the dynamic environment. Int J Adv Manuf Technol 13:227–232. https://doi.org/10.1007/BF01305874
11. Yang K-K, Sum C-C (1997) An evaluation of due date, resource allocation, project release, and activity scheduling rules in a multiproject environment. Eur J Oper Res 103:139–154. https://doi.org/10.1016/S0377-2217(96)00266-4
12. Ebadian M, Rabbani M, Torabi SA, Jolai F (2009) Hierarchical production planning and scheduling in make-to-order environments: reaching short and reliable delivery dates. Int J Prod Res 47:5761–5789. https://doi.org/10.1080/00207540802010799

13. Hicks C, Song DP, Earl CF (2007) Dynamic scheduling for complex engineer-to-order products. Int J Prod Res 45:3477–3503. https://doi.org/10.1080/00207540600767772
14. Elloumi S, Fortemps P, Loukil T (2017) Multi-objective algorithms to multi-mode resource-constrained projects under mode change disruption. Comput Ind Eng 106:161–173. https://doi.org/10.1016/j.cie.2017.01.029
15. Jiang C, Xi JT (2019) Dynamic scheduling in the engineer-to-order (ETO) assembly process by the combined immune algorithm and simulated annealing method. Adv Prod Eng Manage 14:271–283. https://doi.org/10.14743/apem2019.3.327
16. Satic U, Jacko P, Kirkbride C (2022) Performance evaluation of scheduling policies for the dynamic and stochastic resource-constrained multi-project scheduling problem. Int J Prod Res 60:1411–1423. https://doi.org/10.1080/00207543.2020.1857450
17. Xu J, Bai S (2023) A reactive scheduling approach for the resource-constrained project scheduling problem with dynamic resource disruption. Kybernetes 53:2007–2028. https://doi.org/10.1108/K-09-2022-1339
18. Morton TE, Pentico DE (1993) Heuristic Scheduling Systems: With Applications to Production Systems and Project Management. Wiley, New York
19. Dunning JH, Lubin M (2017) Jump: A modeling language for mathematical optimization. SIAM Rev 59:295–320. https://doi.org/10.1137/15M1020575
20. Huangfu Q, Hall JAJ (2018) Parallelizing the dual revised simplex method. Math Program Comput 10:119–142. https://doi.org/10.1007/s12532-017-0130-5

Footnotes

1. The concept Master Production Scheduling (MPS) is commonly used in related contexts.

Figures

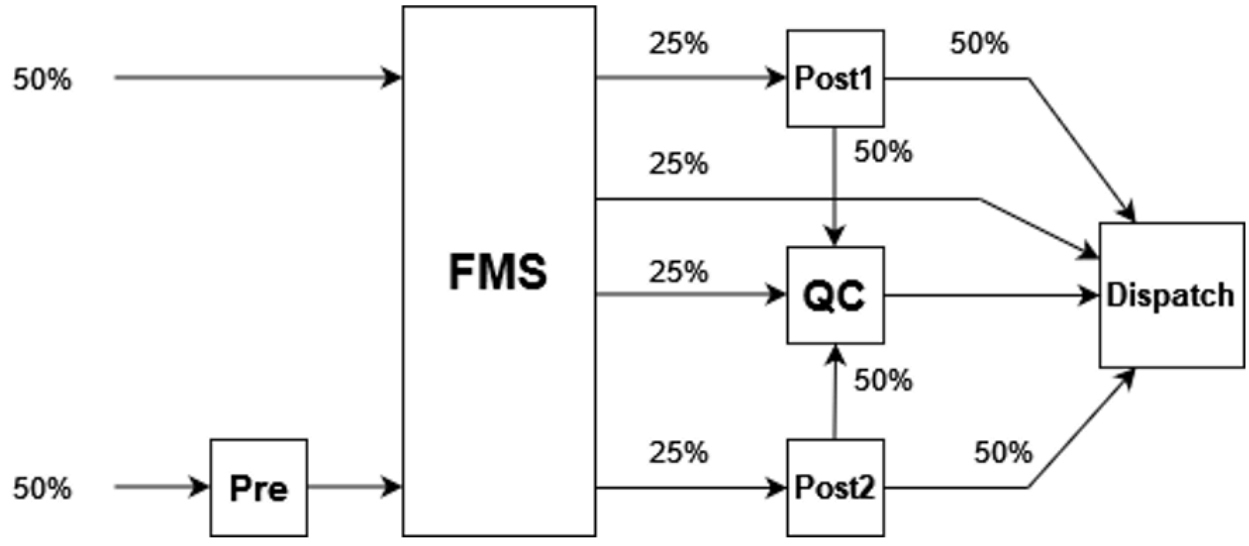


Figure 1

FMS-centered manufacturing system. Percentages indicate the proportion of orders traversing through a given edge

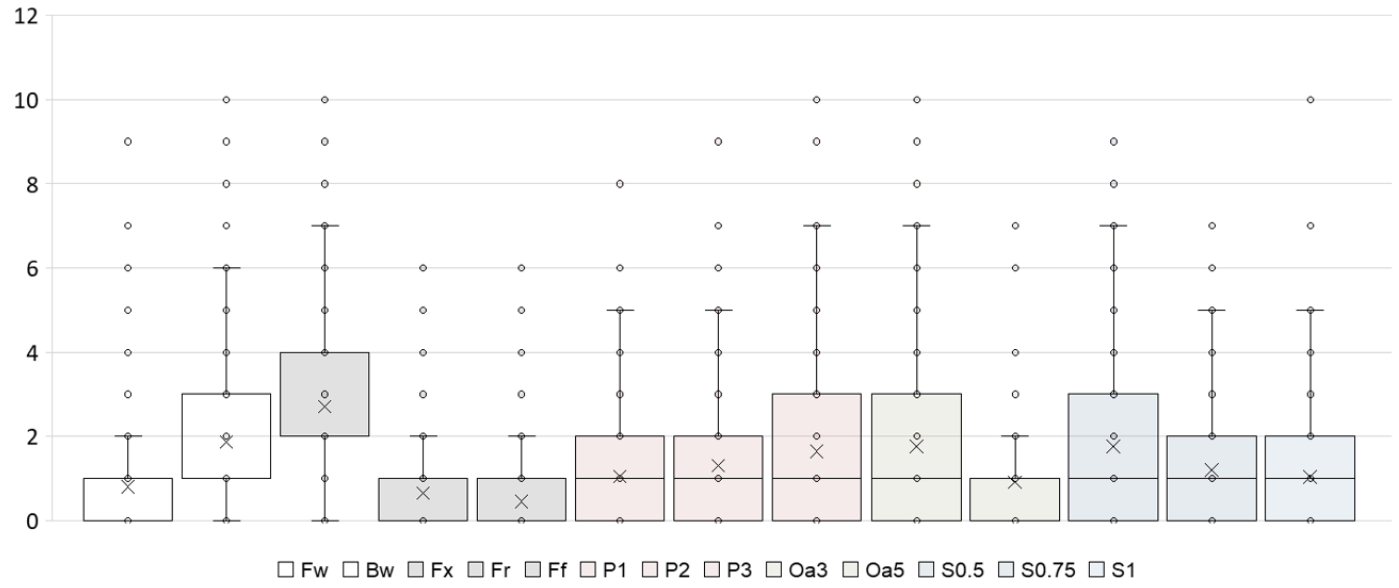


Figure 2

Average maximum tardiness days of orders and associated variability for all different value combinations containing the different independent variables from left to right: Fw = forward planning, Bw = backward planning, Fx = fixed timing and even workload distribution, Fr = free timing and even workload distribution, Ff = free timing and free workload distribution, P1 = planning every day, P2 = planning every other day, P3 = planning every 3rd day, Oa3 = order arrival rate 3/day, Oa5 = order arrival rate 5/day, S0.5 = due date slack coefficient 0.5, S0.75 = due date slack coefficient 0.75, S1 = due date slack coefficient 1

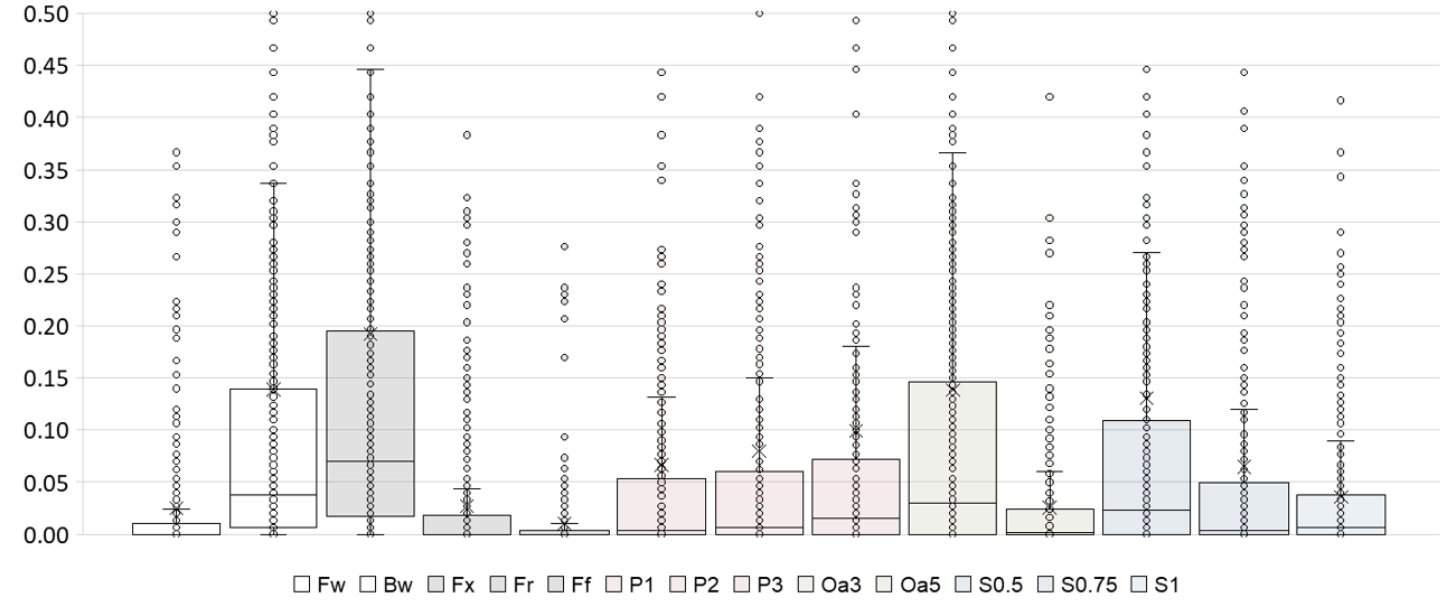


Figure 3

Average total tardiness days per order and associated variability for all different value combinations containing the named independent variables, see Fig. 2 for legend descriptions. The graph is vertically truncated for clarity; highest outlier values are 2.29

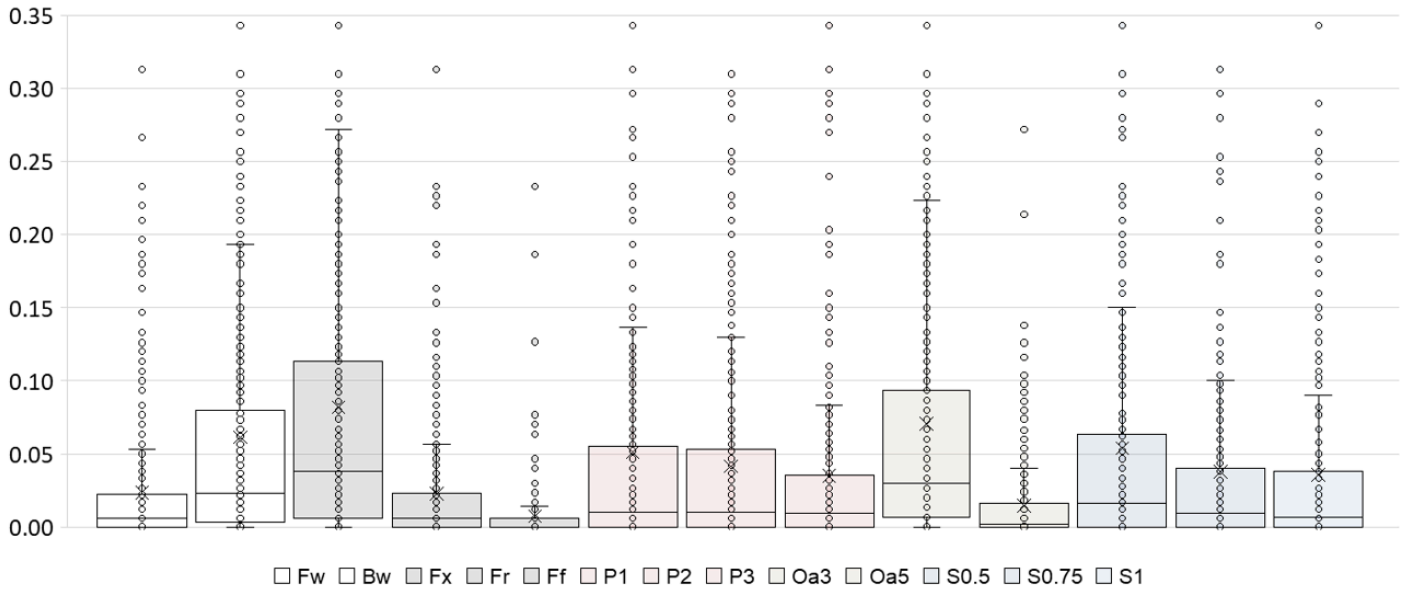


Figure 4

Average total activity reschedulings per order and associated variability for all different value combinations containing the named independent variables, see Fig. 2 for legend descriptions. The graph is vertically truncated for clarity; highest outlier values are 0.767

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx