Anticipation of lead time performance in supply chain operations planning

Citation for published version (APA):

Document status and date:
Published: 01/01/2009

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Anticipation of lead time performance in Supply Chain Operations Planning

Michiel Jansen *, Ton G. de Kok, Jan C. Fransoo

School of Industrial Engineering, Technische Universiteit Eindhoven
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

August 27, 2009

Abstract

Whilst being predominantly used in practice, linear and mixed integer programming models for Supply Chain Operations Planning (SCOP) are not well suited for modeling the relationship between the release of work to a production unit and its output over time. In this paper we propose an approach where the SCOP model is decomposed into a deterministic materials coordination model and a stochastic lead time anticipation model. The sub-models are linked through planned lead times and workload targets. A novel algorithm is presented for rescheduling aggregate workload such that planned lead times can be met by the production unit. The approach that we present is more general than the closely related clearing function concept. It can be applied to multi-item resources without making assumptions on the order of processing, it accurately captures the development of workload over time, and it yields a more reliable production planning. Simulation experiments show that our approach yields reductions in inventory holding costs up to 20% compared to a deterministic materials and resource coordination model.

Keywords: hierarchical production planning; lead time anticipation; supply uncertainty; production smoothing

*Corresponding author, e-mail: m.jansen@tue.nl


1 Introduction

Thanks to advances in information and communication technology, firms nowadays have access to detailed and up-to-date information about the state of their primary process. This enables firms to exercise central control of large and complex supply chain networks. Information sharing across partners in the supply chain, access to point-of-sales data, and integrated sales-and-operations planning, provides valuable advance demand information. On the other hand, a volatile and demanding market requires firms to be highly flexible and provide a reliable supply of a broad range of products, preferably of the shelf. Firms are therefore looking for ways to reduce lead times and inventories to gain competitive advantage but are encumbered by smaller sales quantities and shorter product life cycles for individual products. As a result, the problem of coordinating material flows and release of work-orders to resources has become much more difficult. Decision making is particularly difficult due to many forms of uncertainty in the information upon which decision are based. Different forms of slack (capacity, time, and materials) are created in order to deal with these uncertainties. In this environment, there is a need for reliable dynamic planning methods that efficiently utilize available information on the one hand while taking heed of the uncertainties that are inherent to production and demand processes. The stage for these methods are Advanced Planning Systems (APS). However, present-day APS’s are mainly based on linear and mixed integer programming (LP/MIP) models that are incapable of capturing the uncertainties. In this paper we present a method that augments generic models for supply chain operations planning with the ability to account for uncertainties in the production process.

APS are organized according to a hierarchical production planning (HPP) concept (see Stadtler and Kilger, 2000). The idea of hierarchical decision making is often attributed to Anthony (1965), and Hax and Meal (1975) are the first to present a HPP model. Hierarchical planning systems are structured along the organizational levels of decision making. Decisions made at different levels differ in scope, horizon, and detail of information needed. HPP facilitates human interaction and their decomposition makes the planning problem computationally tractable.

Various frameworks for HPP are proposed in literature. See for example Vollmann et al. (1984); Bitran and Tirupati (1993); Meyr et al. (2000); Hopp and Spearman (2001). Most HPP frameworks separate three levels of control: an aggregate level where long term plans for capacity and volume are decided
on, an intermediate level where the coordination of material flows and assignment of work-orders to resources takes place, and a lower level where the actual execution of production takes place. This methods proposed in this paper concern the intermediate and lower levels.

Two HPP frameworks that are the conceptual basis for our work are proposed by Bertrand et al. (1990, chapter 3) and by De Kok and Fransoo (2003). Both frameworks have a network of production units at the lowest hierarchical level. The production unit (PU) is a self-contained part of the production process where materials are transformed into other materials. Here, transformation may be understood in the most general sense (e.g. assembly, packaging, transport). A key property of a PU is that only the input is controlled by the higher hierarchical level. PUs are decoupled from another by stock-points and may be part of multiple organizations (e.g. supplier, sub-contractor, distributor).

There is a difference in focus of the two frameworks. The framework of Bertrand et al. has a vertical focus. That is, the framework considers an individual production unit. Although there is a separate function for the coordination of material flows, the main objective in this framework is to find the proper loading of the production unit such that reliable and short lead times can be achieved. The framework of De Kok et al. on the other hand, has a horizontal focus. Central to this framework is the coordination of material flows in the production network. Production units are only considered through a constant planned lead time, possible augmented by a deterministic capacity constraint. The method presented in this paper aims at combining both perspectives.

An important design parameter in any HPP concept is the linkage of different levels of control. Poor design of these linkages inevitably leads to inefficiencies and customer dissatisfaction. The terminology provided by Schneeweiss (2003) is instrumental in describing HPP these linkages. Schneeweiss describes planning hierarchies in terms of a top level and a base level. Top and base level are linked through instructions in the top-down direction and responses in the bottom-up direction. The concepts presented by Schneeweiss are more elaborate, but here we only convey the basic ideas applied to HPP. Order releases form the instructions, whereas the response refers to periodic information provided by the PU on its state. Because of processing uncertainties and a phenomenon that is called information asymmetry (see De Kok and Fransoo, 2003), this response may be different than the response that was anticipated by the top level. This yields the concept
of the *anticipation function*. The anticipation function is the representation of the PU as incorporated in the top level model. In reality, the anticipation model can never fully predict the future state of the PU. Therefore there is a need to introduce slack. The amount of slack needed depends on the accuracy of the anticipation function used.

An important parameter in the coordination problem is the planned lead time. The planned lead time prescribes the maximum time between the release of a production order and the instant that the processed order is available for use in downstream stages or for supply to the final customer. It is the responsibility of the PU to dispatch jobs to resources such that orders are processed within the planned lead time. In most practical situations, the PU may only be expected to meet the planned lead time if the amount of work is kept within limits. This requirement translates to additional constraints or additional costs in the objective function of the top level model and are part of the anticipation function.

It was already mentioned that non-linear relations between order release to a PU and its output over time, are not easily captured in LP/MIP models. The anticipation function in these models usually is a pessimistic maximum capacity constraint. As a result, WIP levels in the PU are lower than necessary leading to reduced throughput and long lead times. This problem was recognized by Karmarkar (1989) who proposed the use of a *clearing function* in LP/MIP models for production planning. The clearing function is an anticipation function that relates the total amount of work in the PU to its expected periodic output. Inspired by results from queueing theory, Karmarkar proposes a clearing function \( f(W) = \mu \cdot \frac{W}{W + k} \), where \( \mu \) is the nominal production rate and \( k \) is a parameter determining how fast the nominal production rate is achieved. The concave form of the clearing function allows for a piece-wise linear approximation that is included in the LP/MIP model.

Recently, Selçuk (2007) shows that the periodic output grows much faster with WIP to the nominal production rate than clearing functions based on steady-state analysis of queueing models predict. He proposes the use of expected output conditioned on the initial WIP. Missbauer (2009) presents an approach where an overload phase (workload exceeds period capacity) is distinguished from a transient phase (workload is less than capacity). Clearing in the overloaded phase is approximated by a fluid model and clearing in the transient phase is approximated by an exponential function. A similar transient approach is provided by Riano (2002). Like Missbauer, Rianó models periodic output as a weighted sum of previous inputs. Weights are
estimated from a “transient Little’s law” and the problem is solved in an iterative fashion. Multi-item clearing functions are presented by Hwang and Uzsoy (2005), and Asmundsson et al. (2006). To disaggregate the expected output quantity, both approaches use Little’s formula and assume that remaining lead times for all items are equal. A recent overview of literature on clearing functions can be found in Pahl et al. (2007).

All of the above approaches aim to obtain expected periodic output as a function of initial WIP and planned order releases. There are a number of important limitations to this approach (although the transient approaches overcome some of these limitations). Firstly, as Missbauer (2006) notices, in multi-period planning, there is no fixed functional relationship between expected load and expected output over time since higher moments of the load distribution are different for future periods. Secondly, expectation may not be the best measure of output for SCOP models since it implies a substantial probability that materials are unavailable at the time they are required in downstream stages. It is likely that a more robust type of planning allows for lower overall costs for slack. Finally, in a multi-item setting, periodic output depends on the order of processing of items. A-priori assumptions on the order of processing which are necessary for multi-item clearing functions, are problematic since the PU is free to choose the order of processing as long as the planned lead times are respected.

It is not likely that the limitations of clearing functions mentioned in the previous paragraph are easily overcome. In this paper, we therefore consider the use of an alternative anticipation function that considers lead times rather than periodic output. In literature on production planning under uncertain lead times it is generally assumed that lead times have a load independent identical distribution. That is, capacity is assumed to be infinite. In this literature the aim is either to find the optimal order quantities (e.g. Lu et al., 2003), or to find the optimal planned lead time (e.g. Yano, 1987; Ould-Louly and Dolgui, 2004; Tang and Grubbström, 2003). Cakanyildirim et al. (2000) considers random lead times that depend on the order size but are independent from another. The assumption that lead times are independent is justified if orders are placed at external suppliers and together only consume a small part of the supplier capacity. In this paper, we assume that the capacity is finite and dedicated so waiting times are an important part of the lead time.

The aim of our anticipation function that we propose in this paper, is to verify whether a production plan can be realized within the planned lead
times. In contrast to the periodic output perspective of clearing functions, the lead time perspective does not require assumptions about the order of processing. If necessary, the planned workload is smoothed over time such that it becomes planned lead time feasible. The smoothing results in aggregate release targets are fed back into the materials coordination model such that an improved production plan can be generated. In the adjusted plan, planned releases are such that the production units are able to reliably process them within the planned lead time. As a result, tardiness of orders is reduced, allowing for a reduction of the safety stocks. A simulation study shows that this approach results in a reduction of total inventory holding cost up to 20% compared with an approach where a deterministic maximum capacity is assumed.

The remainder of this paper is structured as follows. First we describe the basic idea behind the lead time anticipation approach. We then postulate the LP formulation for the materials coordination model. Next, we derive an approximate distribution of the workload at the start of each period. We describe a novel algorithm for rescheduling the workload in a PU such that the workload schedule is lead time feasible. The previous two steps together form the anticipation model. We then propose a way to update the materials coordination model with the aggregate release targets returned by the anticipation model. Finally, we discuss the results of a simulation study in which we compare our approach with the purely deterministic approach and the clearing function approach.

2 A lead time anticipation approach

The materials and resource coordination problem in stochastic by nature. However, given the planned lead times and a sales plan, it becomes a deterministic optimization problem that can be solved using LP/MIP techniques. In our lead time anticipation (LTA) approach, the stochastic characteristics of PUs are captured in a separate anticipation model. The SCOP problem is then solved by subsequently generating a production plan in the (deterministic) materials coordination model, evaluating it in the anticipation model, and improving it in the materials coordination model with the feedback of the anticipation model. These steps are also indicated in the HPP concept in Figure 1.

Given the planned lead time, a sales plan (1), and the inventory state of
the production network (2), the *materials coordination model* generates an initial production plan such that customer demand is satisfied with minimal inventories. The production plan contains planned order release quantities for each item and for each period within the planning horizon. The release quantities in each period imply an aggregate workload which is the sum of random processing times for each item. In the LTA model, the resultant schedule of aggregate workload releases (3) is checked against (planned) lead time feasibility. The schedule is lead time feasible if, at all times, the workload can be processed within the planned lead time. Using a novel rescheduling heuristic, workload is scheduled forward to earlier periods wherever the planned lead time cannot be observed. This yields adjusted aggregate release targets (4) that are fed back into the materials coordination model in the form of additional constraints on releases.

The LTA step is applied for each PU individually. As a result of setting aggregate release targets for a PU, material requirements at upstream stages may change. In order to make sure that all changes are taken into account, we consider stages in the network sequentially, starting at the most downstream stage. PUs are combined into stages according to their low-level code (LLC) $v_k$. The resource LLC is the maximum of the LLC's of items produced on that resource: $v_k = \max_{i \in U_k} \{v_i\}$. The item LLC $v_i$ is defined as usual:

$$v_i := \begin{cases} 0, & \text{if } N_i^S = \emptyset \\ 1 + \max_{j \in N_i^S} \{v_j\}, & \text{otherwise} \end{cases}$$
Figure 2: conceptual LTA approach

where $\mathcal{N}_i^S$ is the set of successors of i.

The LTA approach is summarized in Figure 2. The cycle is performed once for each stage in the production network. Only then, the final decision is instructed to the PUs (9). In rolling schedule based planning, orders are released only for the first period in the production plan. The SCOP problem is solved again one period later, using up-to-date information on sales and the state of the production network. In the next three sections we discuss the three boxes in Figure 2 in more detail.

2.1 The materials coordination model

Before we continue with the description of the LTA model, we first postulate the basic coordination model. Our formulation is an adaptation of De Kok and Fransoo (2003). In this adaptation we explicitly model work in process which is only implicitly included in the formulation of De Kok et al. Other examples of LP/MIP models for production planning can for example be found in Tempelmeier (2006); Pochet and Wolsey (2006). The notation used in the model formulation can be found in Table 1. Constraints enumerate over all items $i$ or resources $k$, and over all periods $t = 0, \ldots, H - 1$.

Coordination model

\[
\min \sum_{t=1}^{H} \sum_{i \in \mathcal{N}} c_i^k I_i^+(t) + c_i^k I_i^-(t)
\]
Variable | Description
--- | ---
$H$ | Planning horizon
$L$ | Planned lead time
$D_i(t)$ | Demand for item $i$ arriving in period $t$
$I_i^+(t)$ | Physical stock of item $i$ at the start of period $t$, before receipt of $R_i(t-L)$
$I_i^-(t)$ | Backlog of item $i$ at the start of period $t$, before receipt of $R_i(t-L)$
$R_i(t)$ | Release quantity of item $i$ at the start of period $t$
$X_i(t)$ | Quantity processed of item $i$ in period $t$
$W_i(t)$ | Work in process (WIP) for item $i$, at the start of period $t$, after receipts but before releases
$N$ | Set of all items
$U_k$ | Set of items produced on resource $k$
$CAP_k$ | Periodic capacity of resource $k$
$u_i$ | Capacity consumption for processing one unit of item $i$ on its resource
$c_i^h$ | Holding cost per period for item $i$
$c_i^b$ | Backordering penalty cost per period for items $i$

Table 1: list of notation

Subject to:

(2) \[
I_i^+(t+1) - I_i^-(t+1) = I_i^+(t) - I_i^-(t)
\]
\[-\sum_{j \in N} a_{ij} R_j(t) - D_i(t) + R_i(t-L) + P_i(t),
\]
where $R_i(s) := 0$ for all $s < 0$.

(3) \[
I_i^-(t+1) - I_i^+(t) \leq D_i(t),
\]

(4) \[
W_i(t+1) = W_i(t) + R_i(t) - X_i(t),
\]

(5) \[
\sum_{i \in U_k} u_i X_i(t) \leq CAP_k,
\]

(6) \[
W_i(t) + R_i(t) \leq \sum_{s=t}^{t+L-1} X_i(s),
\]

The objective (1) is to minimize the sum of holding and backlog penalty costs. The inventory balance constraint (2) links inventory from subsequent periods in the horizon together and models the product BOM structure. Constraint (3) restricts backlogging to external demand only (for a discussion on this topic see De Kok et al.). Constraint (4) captures the WIP development in different periods. Constraint (5) limits periodic production to the nominal production rate or production capacity. Finally, constraint (6) restricts releases to what can be produced with the nominal capacity within the planned lead time.

Note that constraints (4)-(6) are the anticipation function for the deterministic materials and resource coordination model. In the LTA approach, these constraints are replaced by the anticipation model and are therefore
omitted in the materials coordination model. Without constraints (4)-(6) the materials coordination model resembles the well-known Materials Requirements Planning (MRP) (Orlicky, 1975). However, it is essentially different due to constraint 3, that ensures allocation of shortages is dealt with by the model itself rather than being left to the planner.

2.2 The anticipation model

The objective of the anticipation model is to ensure that the production plan is (planned) lead time feasible. That is, releases must be such that, at all times, WIP can be processed within the planned lead time. Let $V_t$ be the sum of random processing times of the WIP at the start of period $t$, just after releases. We will call $V_t$ the aggregate workload. We furthermore introduce the PU service level $\phi$ and a function $L^\phi$ returning the $\phi^{th}$ percentile of the distribution of $V_t$. The lead time feasibility is defined as:

\begin{equation}
L^\phi(V_t) := \min \{ l : \mathbb{P} \{ V_t \leq l \} \geq \phi \} \leq L, \quad t = 0, \ldots, H - 1
\end{equation}

$V_t$ depends not only on $\{ R_i(t) \}_{i \in U_k}$ but also on any residual workload not processed in the previous period. Without loss of generality, assume that time is scaled such that the total available processing time is one planning period. Let us define the residual workload variable $V^+_t := \max \{ 0, V_t - 1 \}$, and let the variable $B_t$ denote the additional workload caused by releases $\{ R_i(t) \}_{i \in U_k}$. Then, the development of workload over time is described by:

\begin{equation}
V_t = V^+_{t-1} + B_t
\end{equation}

We assume that processing times are exponentially distributed with rates $\mu_i := CAP_k / u_i$. $B_t$, the aggregate workload for releases in period $t$, is therefore a convolution of Erlang random variables. We approximate $B_t$ by a Gamma distribution. $V^+_{t-1}$ is the residual processing time consisting of the cumulative processing times of all waiting items plus the remaining processing time of the item currently in service. Assuming that, after releases, WIP consists of at least one job in all periods, the sum $V_t := V^+_{t-1} + B_t$ also is approximately Gamma distributed. Note that, if this assumption is not satisfied, constraint (7) is satisfied by definition. Since $V^+_{t-1}$ and $B_t$ are independent, the distribution of $V_t$ is given by its mean and variance:

\[
\mathbb{E}[V_t] = \mathbb{E}[V^+_{t-1}] + \mathbb{E}[B_t] =: \frac{k}{\mu} \\
\text{Var}[V_t] = \text{Var}[V^+_{t-1}] + \text{Var}[B_t] =: \frac{k}{\mu^2}
\]
, where $\mathbb{E}[V^+_t] := 0$ and $\text{Var}[V^+_t] := 0$. Let $F_k$ be CDF of $V_t$, and let $\tilde{F}_k := 1 - F_k$. The mean and variance of $V^+_t$ are (the derivation can be found in the appendix A):

$$
\mathbb{E}[V^+_t] = \frac{k}{\mu} \tilde{F}_{k+1}(1) - \tilde{F}_k(1)
$$

(9)

$$
\text{Var}[V^+_t] = \frac{(k+1) k}{\mu^2} F_{k+2}(1) - \frac{2 k}{\mu} F_{k+1}(1) + F_k(1) - \mathbb{E}[V^+]^2
$$

By means of this recursive Gamma approximation of $V_t$, we are able to evaluate feasibility constraint (7). If this constraint is violated in a period, the production plan must be adjusted. Since delaying the release of work to later periods clearly is not an option (the solution from the deterministic model is an optimistic solution in the sense that infinite capacity is assumed), only forward rescheduling of releases is allowed. Finding the optimal adjustments for the schedule is involved. Through the recursion in equation 8, a rescheduling decision for one period non-linearly affects workload in all subsequent periods. Furthermore, a rescheduling decision and requires making assumptions about the specific items being rescheduled. Rather than attempting to develop an exact optimization algorithm, we propose a heuristic approach based on the assumption that mean and variance are rescheduled proportionally. That is, let $0 \leq p < 1$ the a scalar indicating the proportion of work that is rescheduled from period $t$ to period $t - 1$, then we define the operation $B_t \rightarrow pB_t, B_t > 0$ such that $pB_t \sim \text{Gamma}$ with $\mathbb{E}[pB_t] = p\mathbb{E}[B_t]$ and $\text{Var}[pB_t] = p\text{Var}[B_t]$. Furthermore we define the functions

$$
\tilde{V}_{t-1}(p) = V_{t-1} + p B_t, \quad \tilde{V}_t(p) = \tilde{V}_{t-1}^+(p) + (1 - p)B_t
$$

We make the following two observations:

1. **(Monotonicity)** $L^\phi(\tilde{V}_t(p))$ is non-increasing in the rescheduling proportion $p$.

   \begin{proof}
   Let $0 \leq \tilde{\rho} \leq \tilde{\rho} \leq 1$, $\tilde{V}_t := \max\{V_{t-1} + \tilde{\rho} B_t - 1, 0\} + (1 - \tilde{\rho}) B_t$, and $\tilde{V}_t := \max\{V_{t-1} + \tilde{\rho} B_t - 1, 0\} + (1 - \tilde{\rho}) B_t$. The result follows directly from the fact that for all possible realizations $V_{t-1}, B_t$ we have that $\tilde{V}_t \leq \tilde{V}_t$:
   \begin{align*}
   \tilde{V}_t - \tilde{V}_t &= \max\{V_{t-1} + \tilde{\rho} B_t - 1, 0\} + (1 - \tilde{\rho}) B_t \\
   &\quad - \max\{V_{t-1} + \tilde{\rho} B_t - 1, 0\} + (1 - \tilde{\rho}) B_t \\
   &= \max\{V_{t-1} + \tilde{\rho} B_t - 1 - \tilde{\rho} B_t, -\tilde{\rho} B_t\} \\
   &\quad - \max\{V_{t-1} + \tilde{\rho} B_t - 1 - \tilde{\rho} B_t, -\tilde{\rho} B_t\} \\
   &= \max\{0, -V_{t-1} + 1 - \tilde{\rho} B_t\} - \max\{0, -V_{t-1} + 1 - \tilde{\rho} B_t\}
   \end{align*}
   \end{proof}
Since $-\bar{p} B_t \leq -\bar{p} B_t$, we have $\bar{V}_t - \tilde{V}_t \leq 0$.

2. (Bounded residual work) Residual workload increases if we reschedule a proportion of work from period $t$ to period $t-1$. However, the increase in residual workload $\tilde{V}_t^{+}$ is bounded from above because $\tilde{V}_{t-1}^{+}$ is bounded by constraint (7). Assuming that initially $L^\phi(\tilde{V}_{t-1}^{+}) < L$, $\tilde{V}_{t-1}^{+}$ will first (probabilistically) increase with $p$ up to the point where constraint (7) is tight. After that point, work from period $t-1$ will be rescheduled to earlier periods such that constraint (7) remains satisfied with equality, and $\tilde{V}_{t-1}^{+}$ approximately remains unchanged with a further increase of $p$. (There is no unique distribution of $\tilde{V}_{t-1}^{+}$ such that constraint (7) is satisfied with equality.)

Our workload adjustment heuristic is schematically shown in Figure 3. This figure shows the total workload ($\tilde{V}_t$) just after new releases as a function of the proportion of workload $p$ that is rescheduled to the previous period. Total workload is the sum of residual workload ($\tilde{V}_{t-1}^{+}$) and released workload ($\tilde{B}_t$). Point ($p_1, v_1$) represents the point where constraint (7) is tight for the previous period, that is $L^\phi(\tilde{V}_{t-1}^{+}) = L$. Beyond this point, residual workload increases no further and therefore total workload decreases faster with $p$.

In the following, bold small letters (e.g. $v_t$ are $b_t$) denote vectors with elements mean and variance defining the Gamma distribution of the corresponding random variables with capital letters (e.g. $V_t$ and $B_t$). Furthermore,
\( \Upsilon : \mathbb{R}^2 \to \mathbb{R}^2 \) is the function mapping workload on residual workload (see equation (9)), and \( L^\phi : \mathbb{R}^2 \to \mathbb{R} \) is defined similar as in (7) but now with the random variable in its argument replaced by its defining mean-variance vector.

The algorithm proceeds as follows. First, the total and residual workload are calculated for each period. Starting from the latest period within the horizon and working backward, workload is rescheduled to previous periods whenever constraint (7) is not satisfied. The point \((p_1, v_1)\) is found as follows. The proportion \( p \) and workload \( \hat{v}_{t-1} = v_{t-1} + p \cdot x \) at which (7) is satisfied with equality, are determined using simple line search. Here, \( x \) is the search direction. If \( p \) is negative, then the search direction is \( x := v_{t-1} \) (i.e. existing workload in period \( t - 1 \) must be reduced in order to meet the constraint). If \( p \) is positive, then the search direction is \( x := b_t \) representing that work in period \( t - 1 \) will be increased with work from period \( t \). We set \( p_1 := \max\{0, p\} \) and \( v_1 := \Upsilon(\hat{v}_{t-1}) + (1 - p_1) \cdot b_t \). Next, we establish whether the optimal proportion \( p^* \) is smaller or greater than \( p_1 \). If \( L^\phi(v_1) \leq L \), the optimal proportion is \( p^* \leq p_1 \) and is found in region I in Figure 3. If \( L^\phi(v_1) > L \), the optimal proportion is \( p^* > p_1 \), which is found in region II in Figure 3. The curves in regions I and II are described by:

\[
\mathcal{L}(p) := \begin{cases} 
(I) & L^\phi(\Upsilon(v_{t-1} + p \cdot b_t) + (1 - p) \cdot b_t), & \text{if } 0 \leq p \leq p_1 \\
(II) & L^\phi(\Upsilon(\hat{v}_{t-1}) + (1 - p) \cdot b_t), & \text{if } p_1 < p \leq 1 
\end{cases}
\]

Both functions are decreasing in \( p \) so we use a simple bisection search to find

\[
p^* = \min \{ p : \mathcal{L}(p) \leq L, 0 \leq p \leq 1 \}
\]

The workload smoothing algorithm described here can be found in appendix B. The result of the algorithm is a schedule \( \{\tilde{b}_t\}_{0 \leq t < H} \) that satisfies the lead time feasibility constraint (7) in all periods. However, the original plan may imply a workload for which no lead time feasible plan is possible. In that case, there is an overflow amount of workload which is given by \( \tilde{b}_{t-1} \).

### 2.2.1 Overloading

The existence of any overflow workload indicates that no planned lead time feasible plan is possible that satisfies all dependent and independent requirements. It may be decided that the overflow work should be released to the PU nonetheless such that total backlog costs are kept at a minimum. We refer to
the release of any overflow workload as overloading of the PU. From the literature on clearing functions we learn that overloading of the PU results in a higher expected output, and thus a faster reduction of the backlog. However, overloading also leads to high WIP levels and tardiness of orders (even though the items are available no later than in the original plan). Particularly in convergent network topologies, tardiness may lead to synchronization stocks in downstream stages\(^1\). The alternative strategy is therefore to discard the overflow workload altogether. We refer to this strategy as the conservative strategy and the strategy with overloading as the overloading strategy.

In busy situations, the overflow \((\bar{b}_\cdot)\) may be high. It seems to make no sense to overload the PU directly with a large amount of work. Such a strategy would lead to large and costly WIP levels without significant reduction of the backlog. In a single item setting, the optimal loading of the PU is easily determined. The objective is to minimize the sum of WIP carrying cost and backlog cost.

\[
\min c^w W - c^b \mathbb{E}[X|W]
\]

, where \(W\) is the maximum WIP in the PU and \(\mathbb{E}[X|W]\) is the clearing function of Selçuk (2007). Since \(\mathbb{E}[X|W]\) is concave in \(W\), the cost function is convex and the first order condition is

\[
P\{Q \leq W^\ast\} = 1 - \frac{c^w}{c^b}
\]

, where \(P\{Q \leq x\} := P\{X \leq x|W = \infty\}\). In our multi-item setting the output optimal loading quantity is less easily determined. We define \(f := \frac{c^w}{c^b}\), and rewrite the cost function in a workload equivalent form:

\[
\min_{p, 0 \leq p \leq 1} \left[ f \mathbb{E}[p \cdot V] - \mathbb{E}[\min\{L, p \cdot V\}] \right]
\]

Here, we take into account that the WIP does not leave the PU until the end of the planned lead time. Since the minimum function is concave, this cost function is convex as well. Let the total workload for a period be \(V \sim \text{Gamma}(k, \mu)\), let \(F_k(\cdot)\) be its cdf, and let \(\bar{F}_k(\cdot) := (1 - F_k(\cdot))\). The right part of the cost function is:

\[
\mathbb{E}[\min\{L, V\}] = \int_0^L v dF_k(v) + L F_k(L) = \frac{k}{\mu} F_{k+1}(L) + L F_k(L)
\]

\(^1\)Synchronization stocks are stocks of other materials caused by delayed production due to tardiness of the material.
and the cost function becomes:

\[
(10) \quad \min_{0 \leq p \leq 1} \left[ \frac{p_k}{\mu} \left( f - F_{p,k+1}(L) \right) + L \bar{F}_{p,k}(L) \right]
\]

The fraction \( f \) remains to be determined. Backlog and WIP costs are not specified for workload. Inspired by a well-known relation between fill-rate, backordering cost, and holding cost (see Silver et al., 1998, chapter 7), we set the fraction \( f \):

\[
f = \frac{\phi}{1 - \phi}
\]

The optimal \( p^* \) for period 0 is found though a convex optimization routine. If \( p^* \) is less than 1, the remaining workload \((1 - p^*)V\) is scheduled backward to period 1. Next, the optimal overloading is determined for period 1, and so on until \( p^* = 1 \).

### 2.3 Updating the materials coordination model

If the original production plan is not planned lead time feasible, the aggregate release targets returned by the anticipation model are used to generate additional constraints in the materials coordination model. Two issues that need to be addressed are **disaggregation** and **material feasibility**. The disaggregation issue refers to the fact that the release targets returned by the anticipation model specify only the aggregate amount of work that should be released. No specification is given as to what item types should be produced. In a multi-item setting, constraints corresponding to these aggregate release targets, should be formulated such that orders are released for items that are actually required. The material feasibility issue arises from the fact that the scope of the anticipation model is an individual PU. Material dependencies in the production network are not taken into account. It may not be possible to meet the aggregate release targets because material required for production cannot be made available timely. If the aggregate release targets are rigid, this leads to backlogging of dependent demand for these materials. However, due to constraint (3) in the materials coordination model, this is not allowed and leads to an infeasible coordination problem.

The aggregate release targets that are used for the generation of additional release constraints in the materials coordination model are expressed in mean processing time. In the following sub-sections, the non bold-faced letters \( b_k(t) \)
and $\tilde{b}_k(t)$ denote the mean processing time for orders released in the original and the adjusted plan respectively. Similarly, the adjusted planned releases are denoted by $R_i(t)$.

### 2.3.1 Material feasibility

The anticipation model returns aggregate release targets that, when adhered to, ensure that planned lead times are met. However, the scope of the anticipation model is a PU in isolation. Material dependencies (specified by the bill-of-materials) in the production network are not taken into account. Provided there is no overflow workload ($\tilde{b}_k(-1) = 0$), the anticipation model is a forward scheduling algorithm. Assuming that value is added to the product in every stage, echelon stock of an item is more expensive in a downstream stage than in an upstream stage. Therefore, downstream production plans are not affected by forward scheduling. Upstream releases clearly are affected since forward scheduling causes dependent requirements to move forward in time as well. Since backlogging of dependent requirements is not allowed (constraint 3), increased dependent demand must either be satisfied from inventory that is physically available, or from increased production. If neither is possible, a "hard" aggregate release target yields an infeasible coordination problem. In order to avoid this infeasibility, deviations from the aggregate release targets must be admitted at a penalty cost. Introducing a new variable $O_k(t)$ for the amount of deviation from plan, we get the following constraints:

\begin{equation}
O_k(t) = \sum_{s=0}^{t} \tilde{b}_k(s) - \sum_{j \in U_k} \frac{u_j}{\text{CAP}_k} \left( W_j(0) + \sum_{s=0}^{t} \tilde{R}_j(s) \right), \\
O_k(t) \geq 0, \quad t = 0, \ldots, T - 1
\end{equation}

If there is an overflow workload, the aggregate release targets may imply a downward adjustment of cumulative releases. That is:

$$\sum_{s=0}^{t} \tilde{b}_k(s) < \sum_{j \in U_k} \frac{u_j}{\text{CAP}_k} \left( W_j(0) + \sum_{s=0}^{t} R_j(s) \right), \quad \text{for some } t \geq 0$$

In this case, constraint (11) forces a downward adjustment of releases. This adjustment may also result in downward adjustment of releases in the downstream echelon. As a result, penalty costs in the downstream echelon may increase. Note however, that this cannot lead to an infeasible coordination problem since releases are not bounded below in (11).
We have not yet discussed how a deviation from the aggregate release targets should be penalized. The penalty costs assigned to a deviation from the plan may influence the allocation of capacity shortages. In this paper, we take a simple approach and set these penalty cost equal to the backlog cost of an item. For this reason we make the deviation variable dependent on the item (instead of the resource).

### 2.3.2 Disaggregation of aggregate release targets

Specifying a aggregate release target may lead to planned orders for items for which there is no requirement. This situation occurs if materials used in production are not timely available, and penalty costs can be avoided by producing another item that is not required. We would like to limit planned production to those items for which there is a requirement. These quantities are given by the original production plan.

In the context of the issue described in the previous paragraph, we introduce the notion of *delta blocks*. Delta blocks are sets of subsequent planning periods in which there is a positive cumulative deviation of the released workload in the original plan \( \{b_t\} \) from the target \( \{\bar{b}_t\} \). Let \( \delta(t) := \sum_{s=0}^{t} (\bar{b}_s - b_s) \), \( \delta(-1) := 0 \), we define the set of delta blocks \( \Delta \):

\[
\Delta := \left\{ (t_l, t_u) : \delta(t_l - 1) \leq 0, \delta(t_u) \leq 0, \delta(t) > 0, 0 \leq t_l \leq t < t_u < H \right\}
\]

Penalties can only be incurred for periods within these delta blocks. Also we have for any delta block \( (t_l, t_u) \) that

\[
\sum_{s=0}^{t_l-1} \bar{b}_k(s) \leq \sum_{j \in U_k} \frac{u_j}{CAP_k} \left( W_i(0) + \sum_{s=0}^{t_l-1} R_i(s) \right), \quad \text{and}
\sum_{s=0}^{t_u} \bar{b}_k(s) \leq \sum_{j \in U_k} \frac{u_j}{CAP_k} \left( W_i(0) + \sum_{s=0}^{t_u} R_i(s) \right)
\]

That is, the original cumulative planned production is at least equal to the release targets, both just before the delta block, and at the end of the delta block. This implies there are no material shortages just before and just at the end of each delta block. It is therefore sufficient to specify:

\[
\sum_{s=t_l}^{t_u} \tilde{R}_i(t) \leq \sum_{s=t_l}^{t_u} R_i(t), \quad \text{for all } i \in N_k, (t_l, t_u) \in \Delta_k
\]
3 Simulation experiments

In the preceding section, we have presented a method for SCOP where we captured the stochastic behavior of the PU into a separate lead time anticipation model. Given the planned lead times (and the sales plan), the coordination problem becomes deterministic and can be solved using LP/MIP techniques. The separation of the anticipation function allowed us to model the behavior of a PU more accurately than possible in a LP formulation of the integral problem. We presented a new rescheduling algorithm in order to generate lead time feasible production plans. The objective of our lead time anticipation (LTA) approach is to reduce the investment needed in WIP and finished goods inventory (FGI). In this section we verify to what extend our approach achieves its goal using discrete event simulation.

We compare our LTA approach to an approach where throughput is assumed to be deterministic (DT), and the clearing function approach (CF). We furthermore compare the overloading and conservative strategies discussed in section 2.2.1. We refer to the overloading strategy as LTA-OL and to the conservative strategy as LTA-CS. The clearing function used for comparison is the one proposed by Selçuk (2007, chapters 4,5)\textsuperscript{2} The short term clearing function of Selçuk was selected because it proved to be superior to the long term clearing functions used elsewhere.

The statistic of interest in our simulation experiments is the relative total holding cost (THC) difference (positive figures indicate an improvement):

$$\frac{THC_{\text{benchmark}} - THC_{\text{algorithm}}}{THC_{\text{benchmark}}}$$

, where the algorithm may be LTA-CS or LTA-OL and the benchmark may be DT or CF. To be able to compare inventory costs for the selected approaches, fill rates for end-items are equalized at 98\% using the safety stock adjustment procedure of Kohler-Gudum and De Kok (2002).

Motivated by the idea that PUs have a single bottleneck, they are represented in the simulation model by single server queues. Yet, since several other non-critical activities may be required, processed items arrive at the downstream stock point no earlier than a planned lead time after their release. Demand is dynamic and is known over the planning horizon. The fact

\textsuperscript{2}We use a break-point parameter $\epsilon = 0.20$, the value that yielded best performance in the experiments of Selçuk.
that our interest here is in supply uncertainty and not in demand uncertainty, justifies the use of perfect advance demand information.

3.1 Test bed

The simulation program is coded in the Microsoft Visual C# 2008 programming language. The planning problem is solved every period using ILOG CPLEX 11.0. PUs are modeled as single server queues with exponential processing times where orders are processed in a work-conserving FCFS manner. Processed items (or finished work-in-progress) wait in the PU until the end of the lead time when they are shipped to the downstream stock point. Periodic planning decisions are translated into multiple unit size orders per item that are released to the production in random order. Demand is sampled from a Gamma distribution with squared coefficient of variation 0.5. Perfect advance demand information is available to the planning model in the planning horizon. Demand occurs in unit size lots at the end of the period (there may be multiple lots per period), just before orders are shipped from the PUs. Inventory, work-in-progress (WIP), and finished work-in-progress (FWIP) are measured continuously. The planning problem is solved at the beginning of each period after demand and shipments from the previous period have occurred.

The simulation experiment is set-up along the lines of Law (2007). Every simulation run consists of 30 replications. For each replication, a set of random number streams is generated for each period, resource, and demand generator. Common random numbers are used across different planning approaches. Every replication starts with an empty system and has a warm-up of 100 periods. Statistics are collected over 5000 periods.

3.2 Experimental design

In order to make a rigorous comparison of approaches, we vary the parameters of the production network. One important parameter is the network topology. We carefully select a number of simple topologies offering different challenges. Topologies are identified by their number that is given in brackets. The simplest of topologies is the single item single stage topology (1). Here, material coordination plays no real role. The simplest of models where material coordination is required is the single item 2-stage serial system (2). Synchronization of material flows is important in convergent networks (3)
and allocation of materials is important in divergent networks (4). Finally, we select a topology where available production time in the upstream PU has to be allocated between two components that are both required for assembly of the final product in the downstream PU (5). The resultant topologies are depicted in Figure 4.

There are numerous other parameters that may influence the performance of our LTA approach. LTA aims at smoothing order releases if capacity in some future period is tight, resulting in lower tardiness of orders. Consequently, safety stocks for end items may be reduced, leading to a lower overall inventories. This leads us to formulate the following hypothesis:

1. LTA is particularly useful in situations where capacity is tight. We therefore expect that LTA has a greater impact on performance at higher utilization levels.

2. High variation in service times implies a greater risk of idle time and tight capacity. Therefore, we expect that LTA has a greater relative performance at lower processing rates (implying higher variance).

3. Uncertainty in the supply of components hampers production in the whole echelon. For this reason we think that the LTA has a higher relative performance if the bottleneck resource occurs upstream.

4. A longer planning horizon provides more opportunities for anticipation and increases its relative performance.
5. Longer planned lead times provide more opportunities to respond to changes in the workload and a lower coefficient of variation for the cumulative processing time. Consequently, we hypothesize that the benefit of LTA is lower at longer planned lead times.

6. Since customer service levels are fixed, poor lead time reliability needs to be compensated by high safety stocks for end items. Therefore, we expect that LTA has a greater impact on total cost if there is a high added value in the downstream stage.

Table 3.2 shows the experimental design for the simulation experiments. The PU service level refers to the level $\phi$ used in the LTA approach. For topologies 3 - 5, fixed service levels of 0.7 for LTA-CS and 0.9 for LTA-OL are used. For topologies 2-5, the planned lead time is set equal for the upstream and the downstream stage. Production rates are stated for the downstream stage. The production rate of the upstream stage is determined by the ratio of utilization rates. Processing symmetry determines whether processing times on multi-item PUs are equal or different. In the asymmetric processing case, the processing time of one item is twice the processing time of the other item. Utilization rates for topologies 2 - 5 are shown as pairs ./. referring to the upstream and downstream stage respectively. In the asymmetric case, the stage containing two PUs has one with a utilization rate of 75% and another with a utilization rate of 85%.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Topology 1</th>
<th>Topology 2</th>
<th>Topology 3</th>
<th>Topology 4</th>
<th>Topology 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>PU service level</td>
<td>0.7, 0.9</td>
<td>0.7, 0.9</td>
<td>0.7 / 0.9</td>
<td>0.7 / 0.9</td>
<td>0.7 / 0.9</td>
</tr>
<tr>
<td>Planning horizon</td>
<td>10, 20</td>
<td>10, 20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Planned lead time</td>
<td>2, 3</td>
<td>2, 3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Production rate</td>
<td>10, 20</td>
<td>10, 20</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Processing symmetry</td>
<td>n/a</td>
<td>n/a</td>
<td>symmetric, asymmetric</td>
<td>symmetric, asymmetric</td>
<td></td>
</tr>
<tr>
<td>Utilization rate</td>
<td>0.75, 0.85</td>
<td>0.85 / 0.75, 0.75 / 0.75, 0.85 / 0.75, 0.85 / 0.75, 0.75 / 0.85</td>
<td>0.75 / 0.85, 0.75 / 0.85, 0.75 / 0.85, 0.75 / 0.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value added</td>
<td>n/a</td>
<td>37.5%</td>
<td>20%, 50%</td>
<td>20%, 50%</td>
<td>20%, 50%</td>
</tr>
</tbody>
</table>

The results of the simulation experiments are summarized in Tables 3.2, 3.2, 3.2, and 3.2. Wherever the PU service level is not explicitly mentioned,
the values $\phi = 0.7$ for LTA-CS and $\phi = 0.9$ for LTA-OL were used. Topologies 3-5 are multi-item settings where CF cannot be applied. The results of the simulation experiments are discussed in the next section.

Table 3: Simulation results for LTA-OL, topologies 1 and 2

<table>
<thead>
<tr>
<th>Relative improvement</th>
<th>Topology 1</th>
<th>Topology 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DT</td>
<td>CF</td>
</tr>
<tr>
<td>Internal service</td>
<td>0.7</td>
<td>5.57%</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>8.15%</td>
</tr>
<tr>
<td>Horizon</td>
<td>10</td>
<td>6.47%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>9.87%</td>
</tr>
<tr>
<td>Lead time</td>
<td>2</td>
<td>11.06%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5.24%</td>
</tr>
<tr>
<td>Production rate</td>
<td>10</td>
<td>9.41%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>6.80%</td>
</tr>
<tr>
<td>Utilization rate</td>
<td>0.85/0.75</td>
<td>9.14%</td>
</tr>
<tr>
<td></td>
<td>0.75/0.85</td>
<td>6.76%</td>
</tr>
</tbody>
</table>

Results shown are percentage reduction of total holding cost in comparison with deterministic and clearing function approach respectively.

Table 4: Simulation results for LTA-CS, topologies 1 and 2

<table>
<thead>
<tr>
<th>Relative improvement</th>
<th>Topology 1</th>
<th>Topology 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DT</td>
<td>CF</td>
</tr>
<tr>
<td>Internal service</td>
<td>0.7</td>
<td>4.24%</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>2.58%</td>
</tr>
<tr>
<td>Horizon</td>
<td>10</td>
<td>1.77%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>6.71%</td>
</tr>
<tr>
<td>Lead time</td>
<td>2</td>
<td>6.75%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.73%</td>
</tr>
<tr>
<td>Production rate</td>
<td>10</td>
<td>4.40%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>4.08%</td>
</tr>
<tr>
<td>Utilization rate</td>
<td>0.85/0.75</td>
<td>5.11%</td>
</tr>
<tr>
<td></td>
<td>0.75/0.85</td>
<td>3.37%</td>
</tr>
</tbody>
</table>

Results shown are percentage reduction of total holding cost in comparison with deterministic and clearing function approach respectively. A '*' indicates that the result is not significant ($\alpha=0.05$).

3.3 Discussion

Most of the effects that can be observed are as we hypothesized. We will only discuss some remarkable results. Our simulation results clearly show
Table 5: Simulation results for LTA-OL, topologies 3 – 5

<table>
<thead>
<tr>
<th>Relative improvement</th>
<th>Symmetric</th>
<th>Topology 3</th>
<th>Topology 4</th>
<th>Topology 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processing time symmetry</td>
<td>n/a</td>
<td>13.17%</td>
<td>11.82%</td>
<td></td>
</tr>
<tr>
<td>Utilization rate 0.85/0.75</td>
<td>22.72%</td>
<td>17.27%</td>
<td>14.86%</td>
<td></td>
</tr>
<tr>
<td>Utilization rate 0.75/0.85</td>
<td>10.56%</td>
<td>11.01%</td>
<td>10.27%</td>
<td></td>
</tr>
<tr>
<td>Asymmetric</td>
<td>n/a</td>
<td>13.43%</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>Value added 20%</td>
<td>15.06%</td>
<td>13.53%</td>
<td>12.60%</td>
<td></td>
</tr>
<tr>
<td>Value added 50%</td>
<td>15.04%</td>
<td>14.27%</td>
<td>12.53%</td>
<td></td>
</tr>
</tbody>
</table>

Results shown are percentage reduction of total holding cost in comparison with deterministic approach.

Table 6: Simulation results for LTA-CS, topologies 3 – 5

<table>
<thead>
<tr>
<th>Relative improvement</th>
<th>Symmetric</th>
<th>Topology 3</th>
<th>Topology 4</th>
<th>Topology 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processing time symmetry</td>
<td>n/a</td>
<td>3.22%</td>
<td>4.73%</td>
<td></td>
</tr>
<tr>
<td>Utilization Rate 0.85/0.75</td>
<td>12.79%</td>
<td>5.96%</td>
<td>8.48%</td>
<td></td>
</tr>
<tr>
<td>Utilization Rate 0.75/0.85</td>
<td>0.59%</td>
<td>1.57%</td>
<td>2.88%*</td>
<td></td>
</tr>
<tr>
<td>Asymmetric</td>
<td>n/a</td>
<td>3.90%</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>Value Added 20%</td>
<td>3.82%</td>
<td>3.71%</td>
<td>5.80%</td>
<td></td>
</tr>
<tr>
<td>Value Added 50%</td>
<td>3.62%</td>
<td>3.91%</td>
<td>5.57%</td>
<td></td>
</tr>
</tbody>
</table>

Results shown are percentage reduction of total holding cost in comparison with deterministic approach. A '*' indicates that the result is not significant (α=0.05).
that the LTA-OL approach is superior to all other approaches in terms of inventory holding cost. There is not a single scenario where LTA-OL does not outperform DT and CF. The LTA-CS approach still outperforms DT for most combinations of parameters but, especially for topology 1, its relative performance is lower than CF. Releases in LTA-CS are more restricted than in the other approaches. As a consequence it takes more time to recover from a backlog situation and safety stocks need to be higher. This explains why this approach shows better performance at lower PU service level $\phi$. In multi-stage topologies, the benefit of higher lead time reliability in LTA-CS appears to be somewhat greater indicating that lead time reliability may be more important if there are multiple stages. LTA-OL does not restrict releases and yields better relative performance than LTA-CS.

The planning horizon has a particularly large effect on the relative performance of anticipation functions. If little advance demand information is available, there are no opportunities for anticipation. It remains a topic for future research to study the performance of anticipation functions if the advance demand information is unreliable (e.g. peek demand cannot be foreseen).

Another factor that drastically influences relative cost performance is the planned lead time. If planned lead times are long, a higher WIP is kept in the PU. Pooling effects make that the variability of the cumulative service time for the WIP is low. Furthermore, during a long planned lead time there are more opportunities to respond to changes in the workload. As a result, the relative performance of an anticipation function is reduced. Interpreted differently, this result shows that the LTA approach may allow for shorter planned lead times that would otherwise not be practical because of the high uncertainty.

An interesting outcome of the simulation experiment is that relative performance with anticipation functions in the single stage is lower at higher utilization levels. We may explain this effect as follows. A high utilization level implies that there is a limited amount of slack available in each period. Consequently, a surplus of workload in one period needs to be rescheduled to multiple earlier periods which may be impossible if the horizon is short. This idea is confirmed when we look at the interaction effect of utilization level and length of horizon\(^3\). In scenarios with a long horizon, more periods

\(^3\)This can be seen from more detailed simulation results that are not included in this paper.
are available for rescheduling and the difference between low utilization and high utilization is smaller.

In the simulation experiments for topologies 2 - 5, we compare alternative utilization profiles: one with high utilization in the upstream stage, and one with high utilization in the downstream stage. The results show that both the clearing function approach and the anticipation approach yield higher relative performance for the scenarios with a high utilization in the upstream stage. This may be explained by the fact that anticipation functions create slack in the upstream stage allowing for lower safety stocks of the (more expensive) downstream item. The asymmetric utilization profiles for topologies 3 and 4 result in performances that are roughly the average of the performance for the other two profiles.

Finally, it is remarkable to see that the effect of added value is relatively small. It seems that the benefit of anticipation functions is quite insensitive to the cost structure.

4 Summary and conclusions

In this paper we present a method for supply chain operations planning that explicitly takes into account the stochastic nature of production processes. The method combines the focus on the coordination of material flows in the framework of De Kok and Fransoo (2003) with the focus on the loading of production units in the framework of Bertrand et al. (1990). The key feature of our lead time anticipation (LTA) approach is the separation of the planning problem into a deterministic materials coordination model and a stochastic lead time anticipation model.

Given planned lead times and a sales plan, an initial schedule of future planned order releases is generated by the materials coordination model. This production plan is checked against lead time feasibility in the lead time anticipation model. A production plan is lead time feasible if the production unit is able to process the WIP within the planned lead time in all periods. We have presented a novel algorithm for smoothing the workload in case the initial production plan is not lead time feasible. This algorithm produces aggregate release targets that are used to update the materials coordination model, such that a lead time feasible production plan can be generated.

The LTA approach is relatively easy to implement in existing APS systems. No information other than what is commonly available in APS systems
is required.

There are situations where it is not possible to generate a lead time feasible production plan that timely satisfies all demand. We discussed two strategies for dealing with these situations. In the conservative strategy reliable planned lead times of primary concern and no more work is planned than the amount that can be realized reliably. Any overflow workload is discarded. In the overloading strategy, fast reduction of the backlog is more important than meeting the planned lead times, and high workload levels are maintained in the production unit in order to maximize output.

Results from a simulation study show that the overloading strategy is superior to the conservative strategy in terms of total inventory holding cost. The conservative strategy results in lower levels of output in busy periods. Consequently, higher safety stocks for end items need to be maintained. However, the conservative strategy may be improved further by adding a backward pass to the LTA algorithm where overflow workload is planned in the first period where capacity is available. This is a topic for further research.

The simulation study shows that, in comparison to a purely deterministic planning model with a fixed maximum capacity, the LTA approach leads to reductions in inventory holding costs of 5% to 20%. The LTA approach particularly performs well in the following situations:

- Advance demand information is available for future periods
- There is a bottleneck production unit (high utilization) upstream in the production network
- There is variation in the cumulative processing times of orders
- Planned lead times are few periods long

We have also compared the LTA approach with the closely related clearing function approach and found that the overloading strategy always performs better in terms of inventory holding costs. We have not compared the LTA approach to a multi-item clearing function approach, but we believe that our focus on lead times rather than periodic output allows for more accurate modeling, and expect therefore that the performance of the LTA approach is also better here.

The LTA approach presented in this paper offers many opportunities for fine-tuning and improvement. The following topics are, in our opinion, most interesting for future research.

26
• The LTA algorithm presented in this paper leads to a forward scheduling of planned order releases. However, the items that are affected are not required at an earlier time. These items may thus be allowed a longer planned lead time. This yields the idea of dynamically planned lead times. The concept of dynamically planned lead times was already discussed by Selçuk (2007) and it would be interesting to see how it can be applied to the LTA approach.

• We assumed that the internal structure of the PU can be represented by an single server queue. The conceptual approach presented in this paper provides opportunities to apply other results from queueing theory. Additional research in this area is promising and therefore encouraged.

• We have only considered uncertainty in processing times. Other types of supply uncertainty can be considered.

A Surplus distribution of a Gamma r.v.

Let $V$ be a Gamma random variable with parameters $k$ and $\mu$ such that $\mathbb{E}[V] = \frac{k}{\mu}$ and $\text{Var}[V] = \frac{k}{\mu^2}$. We define the surplus distribution of $V$ over $L$ as:

$$\mathbb{P}[V^L + \leq x] := \mathbb{P}[\max\{0, V - L\} \leq x]$$

Let $F_i(\cdot)$ be the cdf of a Gamma distribution with shape parameter $i$ and scale parameter $1/\mu$, and let $\bar{F}_i(\cdot) := 1 - F_i(\cdot)$. We can derive the first two moments of $V^L +$ directly:

$$\mathbb{E}[V^L +] = \int_0^\infty \max\{0, x - L\} \, dF_k(x)$$
$$= \int_0^\infty (x - L) \, dF_k(x)$$
$$= \int_L^\infty x \, dF_k(x) - L \bar{F}_k(L)$$
$$= \int_L^\infty x \frac{e^{-\mu x} (\mu x)^{k-1}}{\Gamma(k)} \, dx - L \bar{F}_k(L)$$
$$= \frac{k}{\mu} \bar{F}_{k+1}(L) - L \bar{F}_k(L)$$

and

$$\mathbb{E}[(V^L +)^2] = \int_0^\infty (x - L)^2 \, dF_k(x)$$
$$= \int_L^\infty x^2 \, dF_k(x) - 2L \int_L^\infty x \, dF_k(x) + L^2 \bar{F}_k(L)$$
$$= \frac{k(k+1)}{\mu^2} \bar{F}_{k+2}(L) - 2L \frac{k}{\mu} \bar{F}_{k+1}(L) + L^2 \bar{F}_k(L)$$

27
B Workload smoothing algorithm

1. Initialize with $t := H - 1, \tilde{v}_t := b_t$, for all $0 \leq t < H$

2. Set $v_0 := b_0$ and for each period $t := 1..T - 1$ calculate $v_t := \Upsilon(v_{t-1}) + b_t$

3. Set $t := T - 1$

4. If $L^\phi(v_t) \leq L$ then goto step 9

5. If $L^\phi(v_{t-1}) > L$ then
   \[ \hat{p} := \max \left\{ p : L^\phi((1 + p) \cdot v_{t-1}) < L, -1 < p < 0 \right\}, \]
   and $\hat{v}_{t-1} := \hat{p} \cdot v_{t-1}$
   else
   \[ \hat{p} := \min \left\{ p : L^\phi(v_{t-1} + p \cdot b_t) \geq L, 0 < p < 1 \right\}, \]
   and $\hat{v}_{t-1} := v_{t-1} + \hat{p} \cdot b_t$

6. Set $p_1 := \max\{0, \hat{p}\}$, $v_1 := \Upsilon(\hat{v}_{t-1}) + (1 - p_1)B_t$.

7. If $L^\phi(v_1) \leq L$ then
   \[ p^* := \min \left\{ p : 0 \leq p < p_1, L^\phi(\Upsilon(v_{t-1} + p \cdot b_t) + (1 - p)B_t) \right\} \]
   else
   \[ p^* := \min \left\{ p : p_1 < p < 1, L^\phi(\Upsilon(\hat{v}_{t-1}) + (1 - p)B_t) \right\} \]

8. Set $\tilde{b}_{t-1} := \tilde{b}_{t-1} + p^* \cdot \tilde{b}_t$, $v_{t-1} := v_{t-1} + p^* \cdot \tilde{b}_t$, and $\tilde{b}_t := (1 - p^*) \cdot \tilde{b}_t$

9. If $t > 1$ then $t := t - 1$, goto 4 else goto 10

10. Set $\bar{p} := \min\{p \geq 0 : L^\phi((1 - p) \cdot \tilde{b}_0) \leq L\}$, $\bar{b}_{-1} := \bar{p} \cdot \tilde{b}_0$, $\tilde{b}_0 := (1 - \bar{p}) \cdot \tilde{b}_0$

11. Stop.

References


Tang, O., Grubbström, R., 2003. The detailed coordination problem in a two-
level assembly system with stochastic lead times. International Journal of
Production Economics 81, 415–429.

Produktionsplanung und-steuerung in Advanced Planning-Systemen, 6th

Vollmann, T., Berry, W., Whybark, D., 1984. Manufacturing Planning and

Yano, C., 1987. Setting planned leadtimes in serial production systems with