A hierarchical approach for capacity coordination in multiple products single-machine production systems with stationary stochastic demands

Jan C. Fransoo a,*,1, V. Sridharan b, J. Will M. Bertrand a

a Graduate School of Industrial Engineering and Management Science, Eindhoven University of Technology, P.O. Box 513, Pauw F16, 5600 MB Eindhoven, Netherlands
b Department of Management, Clemson University, 101 Sirmine Hall, Clemson, SC 29634-1305, USA

Abstract

We introduce a two-level hierarchical planning and scheduling approach for multi-item, single-machine production systems facing stochastic demand. The hierarchical approach extends the existing heuristics to handle situations where demand levels are high compared to the available production capacity. At the top level of the hierarchy the approach deploys a heuristic procedure for coordinating capacity in the medium term. The heuristic produces target production cycle times and target service levels for each of the products. It iteratively allocates capacity to individual products for maximizing the expected profit while meeting the individual product service level targets set by the management. The bottom level of the hierarchy focuses on the short term and aims to realize the target service level while keeping the production cycles stable. At this level, a fixed-sequence order up-to operational scheduling policy which permits a limited amount of flexibility for reacting to short-term demand variations is implemented. Demand which cannot be met from inventory gets lost. We also present and discuss the results of extensive simulation tests under a wide range of environments that serve to demonstrate the superiority of the approach, especially when the demand level is high.

Keywords: Production; Hierarchical planning; Control; ELSP

1. Introduction

This paper presents a planning and scheduling approach developed by the authors for a multiple products single machine production system based on a glass-containers manufacturing company (bottles and jars). The formulation of the planning and control problem reflects the basic characteristics, described below, of this company. Although the characteristics are typical for many firms in process industries, the current literature in scheduling and production planning and control does not address this full set of characteristics.
The manufacturing company's capacity consists of a set of production lines, each producing a specific set of products. Since the interchangeability of products between the lines is limited, intermediate storage is not possible, and the manufacturing time is short, we consider each of the lines as a single machine. Only one product can be produced at a time in a machine. To change production from one product to another, the molds have to be changed and additional set-up activities have to be performed on the glass forming machines. These activities consume a considerable amount of time. We assume that the set-up times are sequence independent.

Production is continuous ("round-the-clock"), and the production lines cannot be stopped. The glass furnace has to remain on temperature and has to produce glass to keep the process under control. During product change-overs, the liquid glass is led along the glass forming machine and returned into the furnace.

The demand level is very high (between 95% and 100% of available capacity), meaning that not all demand can be satisfied from production. Note that some of the available capacity needs to be used for set-up. This leads to a very high level of utilization. The demand is assumed to be stationary and stochastic. Backordering is not allowed; if any customer order cannot be delivered from stock then the order is lost. Furthermore, differences exist between the various products in terms of contribution margin, inventory cost, and set-up cost. Strategic considerations force the firm to achieve and maintain some minimum level of service (in terms of fill rate) for each product. Likewise, stiff competition limits the firm's ability to increase the price of low margin products. Thus, the production planning and control problem can be stated as follows: Maximize profit subject to service level requirements and capacity constraints.

In this paper, we present a two-level planning and control hierarchy for this type of problem. At the top level of the hierarchy, the focus is on medium-term capacity coordination. It specifies which products to produce for how long (in what quantity). This specification is in line with the company's long-term objectives, in terms of profit maximization and achieving individual products' service levels. The bottom level focuses on short-term operational scheduling and control so as to achieve the medium-term targets set at the top level. The performance of the approach is evaluated in terms of relevant total profit and individual products' service levels.

In the next section, we briefly review the academic literature relevant to the problem outlined above. In Section 3, we present a two-level hierarchical model and a formulation of the top level planning problem. Section 4 presents a heuristic for the top level of the model. Section 5 describes the results of simulation experiments conducted to test the performance of the model. Finally, we discuss the hierarchical methodology in the model based on Schneeweisz's framework (1995) and present our conclusions.

2. Literature review

The problem of lot-sizing and scheduling on a single machine has been addressed by many authors in the management science literature. Although much of the previous work focuses on restricted or different versions of the problem described above, we will briefly review this body of work due to its relation to the problem examined in this paper.

The most common version of the problem is the Economic Lot Scheduling Problem (ELSP). Elmaghraby (1978) presents a review of the ELSP and summarizes the basic characteristics of the problem as: deterministic demand rate, fixed production rate (and larger than the demand rate), sequence-independent set-up times and costs, and the exclusion of backlogging demand. The objective is to find a cyclical production schedule such that the sum of set-up costs and inventory holding costs is minimized. Elmaghraby discusses two kinds of approaches to solve the problem: those that are cycle-based and those that are not. Cycle time is then defined as the amount of time between the start of two consecutive production runs of the same product. Bomberger (1966) presented the basic period concept where individual product cycle times are an integer multiple of a
basic period or fundamental cycle time. This is a necessary – though not sufficient – condition for a feasible schedule. Bomberger’s procedure has been improved significantly by Doll and Whybark (1973). They provide an iterative procedure to determine the cycle times. This work was further developed by Haessler (1979) and Park and Yun (1984). Dobson (1987) further elaborates on this work by proposing an approach where the cycle times are fixed but the lot sizes can be varied over time.

The majority of the published research on lot-sizing and scheduling has focused on deterministic problems. Motivated by the lack of effort on the multiple product single machine scheduling problem with uncertain demand, Vergin and Lee (1978) investigate the use of deterministic models in a stochastic environment. They conclude that the cyclical schedules produced by the deterministic models are of little value in a stochastic environment. They surmise that dynamic scheduling rules which take the current inventory position into account should be used when demand is stochastic.

Leachman and Gascon (1988) continue this line of work – studying the problem of lot-sizing and scheduling under stochastic demands. They show – in accordance with Vergin and Lee – that the traditional approaches aimed at deterministic demand situations do not lead to satisfactory results in stochastic situations. Consequently, Leachman and Gascon present a heuristic for situations with a stochastic demand. Their heuristic is a period-based heuristic which attempts, in each review period, to achieve a target cycle time. This target cycle time is calculated based on an economic manufacturing quantity (using an extension of Doll and Whybark’s (1973) procedure). If this target cycle time does not lead to a feasible schedule in the short term (i.e., one or more products run out of stock), then the cycle time is decreased. In this way, production runs can be started earlier than indicated by the target cycle time to increase the (short-term) service level. The total decrease of the operational cycle time is limited by the so-called minimum run-length. The length of a production run can never be shorter than this minimum run-length. One of the essential characteristics of the environment in which the performance of the heuristic is tested is that overtime is allowed (no round-the-clock production). In a later paper Gascon et al. (1992) investigate the heuristic in a production system without any overtime opportunity. The results of these tests indicate that the heuristic performs good if the demand levels are not extremely high. In all cases investigated, their heuristic performs better than some traditional ELSP heuristics (independent EMQ, a Doll and Whybark (1973) based policy, and a Vergin and Lee (1978) based policy).

Gallego (1990) presents a real time scheduling tool for a single machine system producing multiple products facing random demands. He, like us, assumes that the setup times are not sequence dependent. However, unlike us, he permits backorders. In his three step approach, first a (near) optimal target cyclic schedule is computed using constant expected demand rates. Next, a recovery policy is formulated to return to the target schedule after a disruption of the schedule or the planned inventories occurs. Finally, in the third step, appropriate safety levels are defined so as to minimize the long run average cost.

Virtually all published approaches assume that the product demand rates are such that essentially all demand can be satisfied (i.e., the cumulative (total, aggregate) rate of demand is substantially less than the aggregate production rate). However, as described in the previous section, many companies face a cumulative demand rate equal to or greater than the production rate. Therefore we expect that none of the published approaches can be used for the problem addressed in this paper. Fransoo (1992) investigated the case where demand rate exceeds the production rate and compared the behavior (performance) of a variable cycle time policy (e.g., Leachman and Gascon, 1988) and a fixed cycle time policy. His results indicate that when the cumulative demand rate is close to or exceeds the production rate a fixed cycle time policy, in which the cycle times are not allowed to change in the short term, is superior to a variable cycle time policy.

The research presented in this paper repre-
sents a three-fold extension of the published literature on the stochastic lot-sizing and scheduling problem. Specifically, this study extends the scope of the stochastic lot-sizing and scheduling problem in the following ways.

(a) We explicitly include the differences in contribution margin and cost structure as part of the production planning and control problem and focus on long run profit maximization unlike the existing literature, which largely focuses on cost minimization.

(b) Since the aggregate demand rate is high compared to production rate, achieving high level of utilization is extremely important for maximizing profit. Hence, the production system considered in this study operates at a considerably higher level of utilization than that in the published research. Since we assume a lost-sales model of demand, this utilization rate refers to the ratio of customer demand (as opposed to accepted demand) and production capacity.

(c) The control aspect is much more dominant in our problem than it is in the published research. The proposed approach strives to keep the system balanced and stable. It primarily feeds forward to meet the target cycle times instead of using feedback control (cf. Gallego, 1990). As Bourland (1994) states, a stable system is also important if the single machine is part of a larger production system, where the machine’s output is used in other production steps.

3. A two-level hierarchical model

We propose a two-level hierarchical model for solving the problem described earlier in Section 1. Our approach adopts two aspects from previous studies. First, our model consists of a two-tiered hierarchy with a structure similar to the two-step approach found in Leachman and Gascon (1988). Second, we deploy the procedure by Doll and Whybark (1973) in our heuristic to determine the target cycle times at the top level. (See Section 4.)

For clarifying the basic idea of the approach, consider a firm manufacturing three products (A, B, and C) on a single machine. If demand is constant, the inventory position will approximate the pattern shown in Fig. 1, where the horizontal axis is the time axis and the vertical axis represents the inventory quantity. Set-ups are represented by the black rectangles on the horizontal axis.

Suppose the demand rate for product A temporarily increases. If the production department reacts by advancing the start of the next production run of product A, then the cycle time of A reduces, and the production run of product C cannot be completed as planned (see Fig. 2). This is caused by the fact that the set-up for product A has to be performed earlier than planned (the white rectangle in Fig. 2).

All else remaining the same, in the next cycle, product C will run out of stock earlier than planned, since the run length of product C was

![Fig. 1. Regular inventory pattern with constant demand rate.](image1)

![Fig. 2. Inventory pattern if demand exceeds the cycle time capability.](image2)
shortened. This would cause a similar reaction, i.e., reducing the run length of product B to advance the start of the next production run of product C. Ultimately, a chain reaction sets in and as a result of the incidental spike in the demand rate for product A and the shop reaction to shorten the run length of product C, the cycle times of all products are reduced. The reduction in cycle times affects the distribution of the available capacity in such a way that less capacity is available for production and more capacity is spent on setting up. Consequently, even the regular demand level (which is likely to resume after the temporary spike) cannot be met. Restoring the original distribution of available capacity requires an increase of the cycle times (and in corresponding inventory levels). If the production run length of a single product is increased, a part of the extra products produced will be used to cover demand during the increased length of the cycle and the rest is used to restore the inventory level. This will eventually restore the service level again to its original level. However, due to the increased production run length of a single product, production of all other products needs to be postponed. Overall, this situation will lead to a considerable period of time during which an excessive fraction of demand may not be filled.

In fact, in the situation described above short-term interests (flexibility) dominate over long-term interests (total throughput and profit). A policy aimed at short-term results will focus on realizing a high service level on any short-term demand. However, this influences considerably the distribution of capacity. More specifically, it reduces the fraction of capacity available for production. Due to the high demand levels, this will lead to long periods with unfilled demand and influence considerably the long-term profitability of the firm. Therefore, under the assumption that the firm’s objective is to maximize its long-term profit, we hypothesize that it is a sufficient condition for maximizing profit in a highly utilized single-machine production system with large setup times to keep cycle times stable (Fransoo, 1992).

Under a stable cycle times policy, the variance of cycle time for each product is low. This differs from constant cycle times, where the variation equals zero. When the product cycle times are kept stable, it will enable minimizing the amount of capacity expended on set-up and thus increase productive use of capacity. This, in turn, will enhance the coordination between long-term tactical plans and short-term operational decisions. However, a problem might occur, if the actual demand for individual product(s) is consistently different from the average rates used for long-term planning. When the demand level is greater than anticipated and the cycle time is held stable, the excess demand is lost and the service level decreases. Note that in this case there is no loss of productive capacity since run-lengths are not decreased. On average, the length of the production run will equal the order-up-to level minus the expected inventory at the end of the cycle. However, due to demand variations, the production run may be shorter in some instances, and longer in others. When the demand level is less than anticipated and the cycle time is held stable, however, inventory can increase in an uncontrolled fashion. It may be necessary to correct for this, by adjusting the batch size (and, perhaps, the cycle time) if and only if the aggregate inventory rises above a pre-set upper limit. Thus, it is clear that operating under a regime of stable cycle times will reduce cycle time variance and, in turn, will increase productive capacity.

When a single facility committed to producing multiple products is facing a very high level of aggregate demand (close to its capacity), profit maximization is achieved only when the available capacity is allocated judiciously. In fact, in such situations production scheduling concerns two interrelated issues: 1) to distribute capacity between products and 2) to optimize capacity expended in set-up. Clearly both decisions should be based on the long-term objective of profit maximization. In our two-tiered hierarchical model both these issues are addressed, from a long-term perspective, at the top level. At the bottom level the model strives to maintain the integrity of implementation, in the short-term, while being sensitive to day-to-day fluctuations in demand.

The objective at the top level of the hierarchi-
cal model is to determine the values of the system parameters in such a way that the expected relevant profit is maximized, considering the service level requirements. The service level requirements are the management constraints on the service level. When cycle times are stable, the performance of the system is determined by two parameters: cycle times and target inventory levels for each product. If product cycle times are increased, the probability that the available inventory level is insufficient to fill the demand will increase. This will affect both the service levels and the profit. The target inventory level is defined as the level which determines the lot size (i.e., the lot size equals the target inventory level minus the actual inventory level). Therefore, the target inventory level is a parameter. Normally, this level will not be reached, since demand is likely to be filled during the production run. Increasing the target inventory level of a product will not only increase the service level for that individual product (thus increase its contribution to profit) but also lead to an increase in inventory cost.

At the top level, our hierarchy explicitly considers these trade-offs in distributing capacity between products and for set-up so as to maximize the profit. Even though the focus here is on the long term, and the short-term demand dynamics are ignored, it is important to recognize that such dynamics will have an impact on implementation. The short term dynamics are addressed at the bottom level: operational scheduling.

Next, we describe both levels of the hierarchy in more detail in the next two subsections.

### 3.1. Description of the top level: capacity coordination

The purpose of the top level is to set the system parameters (cycle time and target inventory level). These system parameters determine the distribution of the available capacity over productive capacity and set-up capacity. Productive capacity is defined as the fraction of capacity which is available for production; set-up capacity is defined as the fraction of capacity available for setting up. Additionally, the distribution of the productive capacity over the various products is determined. To identify this function of the top level, this level is called Capacity Coordination.

In this section, a nonlinear programming model will be presented for the capacity coordination problem. This model attempts to maximize the long run expected profit subject to management specified minimum service level requirements. It represents the behavior of the single machine under the assumption of (i) stable cycle times (i.e., at the operational level, cycle time variance is low and demand which cannot be met from inventory is lost) and (ii) lot sizing using the target inventory level. The target inventory level of product $i$ is indicated by $I_i^*$. The target cycle time of product $i$ (indicated by $T_i^*$) is the cycle time which results if at each production run the selected product is produced in a lot size which equals the target inventory level minus the inventory level at the start of the run. For each product, the target cycle time is a function of the target inventory level, as is the expected service level. In this paper, we measure the service level as the fill rate, i.e., the portion of demand that is filled out of stock. Given a specific cycle time and a specific target inventory level, the expected fill rate can be determined. The expected fill rate of a product $i$ will be denoted as $\hat{\alpha}_i$.

If we know the expected fill rate, the target inventory level, and the target cycle time, we can determine the expected profit per period, which is the objective function in our model. The expected profit consists of the expected contribution minus the expected inventory holding and set-up cost. The expected contribution is determined by the fraction of demand that is expected to be filled:

$$\hat{\alpha}_i d_i b_i$$

where

- $d_i = $ Average demand for product $i$ (units per period).
- $b_i = $ Contribution margin per unit of $i$.

At this level of aggregation, we estimate the average inventory as the mean of the highest and the lowest expected inventory positions. The highest inventory position is the target inventory
level, reduced by the demand during the production run:

\[ I_i^* (1 - d_i/p_i) \]  

(2)

where

\[ p_i = \text{production rate of product } i \text{ (units per period)}. \]

The lowest inventory position is the inventory level at the end of the cycle (just before the start of a new production run). The expected inventory position at the end of the cycle has been determined using Brown's partial expectation function in the complementary form as has been done in the Appendix (Eq. (A.5)):

\[ \sigma_i T_i^* G \left( \frac{d_i T_i^* - I_i^*}{\sigma_i T_i^*} \right) \]  

(3)

where

\[ \sigma_i = \text{Standard deviation of demand for product } i \text{ per period.} \]
\[ G(\cdot) = \text{Partial expectation value according to Brown (1963).} \]

Note that the \( G(\cdot) \) function is not an expectation operator with a stochastic variable, but a deterministic function expressing the expected quantity short, where the function argument is the given inventory level. At the aggregate level, we assume the demand to be distributed normally, with mean \( d_i \) and standard deviation \( \sigma_i \).

The consequences of this assumption will be further discussed later on, in Section 6.

The last part of the objective function covers the set-up cost. The set-up cost is directly related to the target cycle time (which is proportional to the reciprocal of the set-up frequency). The entire objective function is represented in expression (4):

\[
\sum_{i=1}^{n} \left( \tilde{a}_i d_i b_i - 0.5 \left( I_i^* (1 - d_i/p_i) \right) + \sigma_i T_i^* G \left( \frac{d_i T_i^* - I_i^*}{\sigma_i T_i^*} \right) h_i - u_i \right) \right) - \sum_{j=1}^{n} \left( L_j + c_j \right)
\]

(4)

where

\[ h_i = \text{Inventory holding cost.} \]
\[ u_i = \text{Set-up cost.} \]

There are strong interdependencies between the variables in expression (4). The cycle time and the target inventory level are directly related to the expected fill rate, as has been explained in the previous section. Additionally, strong interdependencies exist between the various products. The cycle time of a product \( i \) is dependent upon the cycle time of all other products \( j \). We will introduce expressions for the target cycle time and the expected fill rate. In these expressions the various interdependencies will be formalized.

First, we derive an expression for the target cycle time. In order to get a set of easily manageable cycle times, we assume that each product is produced according to an integer multiple \( k_i \) of a basic cycle time (analogous to the approach by Doll and Whybark, 1973). Note again that the cycle times in our approach do not lead to a predetermined, deterministic schedule. Instead, they are used as control parameters for the procedure, at the bottom level, which actually determines the operational schedule, given the most recent available information on the inventory position. This justifies using a straightforward cycle time determination, since the characteristics that more complex scheduling approaches consider, are captured in our model at the operational level.

During the cycle time of a product \( i \), the following processes take place:

(i) production of products \( j \) (\( j = 1, \ldots, n \)). Obviously, only if product \( i \) has the longest cycle time of all \( n \) products, all products are produced during the cycle time of product \( i \). If the cycle time of product \( i \) is shorter, then each product \( j \) is produced on average with a frequency \( k_i/k_j \).

(ii) set-up of products \( j \) (\( j = 1, \ldots, n \)). Analogously to production, during the cycle of product \( i \), every product is set up \( k_i/k_j \) times on average.

Given (i) and (ii), the cycle time of product \( i \) can be expressed as

\[
T_i^* = \frac{\sum_{j=1}^{n} \left( k_i/k_j \right) \left( L_j + c_j \right)}{k_i}
\]

(5)
where
\[ c_j = \text{Set-up time for product } j \] (periods per set-up).
\[ L_j = \text{Average length of a production run of product } j. \]

The average production quantity of a product is the difference between the target service level and the average inventory level at the end of the cycle time of this product. The average inventory level at end of the cycle time can be formulated as in expression (3). Consequently, a production run consumes the following duration of time:

\[ L_j = \left( I_j^* - \frac{\sigma_j T_j^*}{\sqrt{\sigma_j^2 T_j^*^2}} \frac{d_j I_j^*}{\sigma_j T_j^*} \right) / p_j. \] (6)

Note that the numerator is the production quantity, obtained as the difference of the target inventory level of product \( j \) and the expected inventory level of product \( j \) which remains, on average, at the end of a cycle time \( T_j^* \) (from expression (3)).

Given Eqs. (5) and (6), the relation between the target cycle time, the target inventory levels and the integer multiples can be expressed in Eq. (7).

\[ T_i^* = k_i \sum_{j=1}^{n} \left( \frac{I_j^*}{k_j p_j} \frac{\sigma_j T_j^*}{\sqrt{\sigma_j^2 T_j^*^2}} \frac{d_j I_j^*}{\sigma_j T_j^*} \right) + \frac{c_j}{k_j}. \] (7)

Both to the time allocated for production runs (from Eq. (5)) and the time allocated for setting up (\( c_j \)), the ratio \( k_i / k_j \) has been added. Note that Eq. (7) controls the allocation of capacity. Furthermore, since \( T_i^* \) appears on both sides of the equation, an iterative procedure is needed to solve it, given the \( k_i \) and \( I_i^* \) values.

Given the target inventory levels, the expected service levels can be determined. Remember that the service level is defined as the fill rate. Note that any demand that is not filled from stock gets lost (no backordering). We define the expected fill rate of a product \( i \) as the fill rate that is expected for product \( i \), given the characteristics of demand (mean and variance), the target cycle times, and the target inventory levels. An expression for the expected fill rate (\( \hat{a}_i \)) is presented in Eq. (8).

\[ \hat{a}_i = 1 - \frac{\sigma_i G \left( I_i^* - \frac{d_i T_i^*}{\sigma_i T_i^*} \right)}{d_i T_i^*.} \] (8)

Eq. (8) shows that the expected fill rate is computed by deducting the fraction of demand that will not be delivered (on average) from 1.

Eqs. (4), (7), and (8) describe the real system behavior in approximate terms. The parameters to be influenced at this level are the target inventory level \( I_i^* \) and the integer multiples \( k_i \). Once these are set, the fundamental cycle time is determined according to Eq. (7). Maximization of expression (4) as a function of \( I_i^* \) and \( k_i^* \) (replace \( T_i^* \) in expression (4) by the RHS of Eq. (7), and \( \hat{a}_i \) in expression (4) by the RHS of Eq. (8)) should result in the approximate optimal setting of the parameters. Obviously, the objective function is very complex. The function is non-linear in its decision variables \( I_i^* \) and \( k_i \), and also the interaction between each of the product cycle times and inventory levels is complex.

Since this model is an approximation, modeling errors are made. In this respect, a number of characteristics of the operational system behavior are not modeled in the approximate model. First, the aggregate model supposes a completely fixed schedule. In reality, however, the cycle times will vary a little bit (due to the fact that the run lengths are not fixed) and the sequence may vary (depending on the sequencing rule which is being used). This will lead to different service levels, inventory levels and set-up costs. Second, the inventory pattern will not behave like a linear function. Finally, at the operational level the orders are accepted on a day-to-day basis, while the approximate model assumes demand on a continuous time scale. However, since detailed modeling of the real system behavior is impossible, we have to rely on an approximate model. Thus we should examine the 'fit' of this model to 'reality'. This especially refers to the relevant profit and the individual product service levels. The quality
of the approximate model in this respect will be referred to as the **internal validity** of the model.

The objective function is maximized subject to a service level constraint. It is important that some management-specified minimum service level is guaranteed for each product. This decision should be taken at the strategic decision level and should be a constraint for the tactical decision described in this section. A minimum service level is realistic from a business point of view. If it did not exist, it might be possible that for some product(s) – especially those with a small contribution margin – no demand will be accepted during the year. In that case, it is unlikely that the product would be part of the product mix of this firm. Although, it may be tempting to increase the price of low margin items, given that the aggregate demand rate exceeds the production rate, often stiff competition precludes this option. The firm is forced to continue to offer low margin items in order to obtain orders for its high margin items. In order to achieve a feasible production schedule, it is necessary that the minimum service level can be met within the capacity constraints. These service level constraints are represented in Eq. (9).

\[ \hat{a}_i \geq \alpha_i^- \]  

where

\[ \alpha_i^- = \text{Predefined minimum service level for product } i. \]

Formally, the capacity coordination problem can now be stated as follows.

**Problem CC**

Maximize \( (4) \) subject to \( (9) \).  

The capacity coordination problem aims to determine the cycle times and the distribution of the available capacity over the product range. The model considers the cost structure of the various products, the capacity they consume, and the demand distribution of each of the products. The model also takes into account the differences in target service levels due to the fact that some \( I_i^* \) are larger than some \( d_i T_i^* \). The aggregate model further assumes that at the bottom level cycle times are kept stable by determining the batch quantity as the difference between the target inventory level and the actual inventory level. Conceivably other operational procedures can be implemented than the one assumed in this aggregate model.

Due to the complexity of the problem, a heuristic is proposed to find a solution. This heuristic will be presented in Section 4.

**3.2. Description of the bottom level: operational scheduling**

The operational scheduling level has to perform two functions, namely production sequencing (which is the next product to be produced?) and production lot-size determination (what is the production quantity?). The objective is not to optimize some function but to control the detailed system behavior such that the expected fill rates are realized, and, consequently, the expected profit is realized. This is achieved by implementing a stable cycle times policy. Therefore, the following operational scheduling procedure has been designed:

(a) The products are produced in a predetermined fixed sequence (obviously taking into account the integer multiples).

(b) A product is produced in the quantity, resulting from the difference between the target inventory level and the actual inventory level at the start of the production run, which enables some reacting to short-term demand variations. Thus:

\[ Q_i = I_i^* - I_i. \]  

In this way, the stability of cycle times will be maintained (Fransoo, 1992), while reactions to short term demand fluctuations are still possible in a limited way (mix variations during a cycle). Note that although the above procedure is simple, it strives to maintain a stable schedule as per the assumption made at the top level in deriving the cycle times and target inventory levels.

**4. Capacity coordination heuristic**

In this Section, we present a heuristic procedure, called CCH, for solving the capacity coordination problem described earlier in Section 3.1.
A rough analysis of the objective function of problem CC indicates that Eq. (3) has a maximum. At some combination of target inventory levels, an increase in inventory would lead to an increase in holding costs which exceeds the increase in expected contribution. At the same combination of inventory levels, a reduction in inventory would lead to a reduction in inventory holding cost which is less than the reduction in expected contribution. The function value around the maximum, however, is expected to be rather flat, since many combinations of the individual target inventory levels may generate about the same result.

Given target inventory levels and $k_i$ values, the objective function value can be easily found. The integer multiples can be determined by minimizing the total cost for a given fill rate. Using these two observations, a heuristic has been constructed. This heuristic is presented below.

The basic idea of the CCH heuristic is to allocate some capacity to a product, then compute the corresponding optimal cycle times. Given the cycle times and the allocated capacity (represented by the target inventory level), the expected profit (which is defined as the gross contribution minus the inventory holding and set-up costs) can be determined. If the allocation of capacity results in a profit increase, some capacity can again be allocated to a product. Each time capacity is allocated, the ratio of productive capacity and set-up capacity is increased. The product to which capacity is allocated is selected based on the expected contribution. If the increase in profit over a number of iterations drops below a predefined critical value, the heuristic stops.

**CCH Heuristic**

**Step 1.** For all $i$: $k_i := 1$.

**Step 2.** For all $i$: Set $I_i^*$, so that $\hat{\alpha}_i = \alpha_i^-$. 

**Step 3.** Based on the actual value of $I_i^*$, compute $T_i^*$, for all $i$, according to Eq. (7).

**Step 4.** Determine the integer multiples according to the IM procedure presented below (steps A–D).

**Step 5.** If any of the multiples have been changed in Step 4, then adjust $I_i^*$, for all $i$, so that all $\hat{\alpha}_i$ remain unchanged.

**Step 6.** Calculate the expected profit according to expression (4). If the increase in profit during the last $\beta$ iterations is less than $\gamma$, then STOP.

**Step 7.** Of all $i$, determine $\ell$, for which

$$b_\ell p_\ell T_\ell^* \left(1 - P \left( \frac{I^*_\ell + 1 - d_\ell T^*_\ell}{\sigma_\ell T^*_\ell} \right) \right)$$

is maximum where $P(\cdot) = \text{Cumulative probability (standardized normal distribution function)}$.

**Step 8.** $I^*_\ell := I^*_\ell + 1$. Go to Step 3.

In Step 1, all $k_i$ are set at 1. This is necessary to be able to calculate the $\hat{\alpha}_i$'s for the initial inventory levels in Step 2.

In Step 2, initial inventory levels are determined such that the minimum requirements set by the management are satisfied. The expected fill rate formula (8) has been deduced from Brown's partial expectation function and has been explained above. According to Brown (1963), the expected quantity short - given a certain inventory level - is the partial expectation value of that inventory level, multiplied by the standard deviation.

In Step 3, the current value of the target inventory level and the integer multiples are used to determine the corresponding target cycle times using Eq. (7). Since $T_i^*$ is on both the left hand side and on the right hand side of the equation, determining the cycle times is not straightforward. In the computer program, an iterative procedure is included, which can start with arbitrary values of $T_i^*$. Consecutive iterations adapt the values of $T_i^*$ until both sides of the equation result in the same value. The starting values of $T_i^*$ determine the number of iterations. The first time that Step 3 is reached, arbitrary values of $T_i^*$ are used. From the next time on, the most recent values of $T_i^*$ are used as starting values.

In Step 4, the integer multiples with the lowest cost are determined using the IM heuristic described below. It adapts the first four steps of Doll and Whybark's heuristic (1973).
IM Heuristic

**Step A.** Determine $T_i$ independently for each product by

$$T_i = \sqrt{2u_i / \left( \hat{\alpha}_i d_i h_i \left( 1 - \hat{\alpha}_i d_i \right) / p_i \right)}.$$  \hspace{1cm} (13)

**Step B.** Select the smallest $T_i$ as the initial estimate of the fundamental cycle time $T$:

$$T = \min(T_i).$$ \hspace{1cm} (14)

**Step C.** Determine the integer multiple $k_i^-$ and $k_i^+$ for each product defined by

$$k_i^- \leq T_i / T \leq k_i^+$$ \hspace{1cm} (15)

where $k_i^- = \text{The next lowest integer multiple.}$

$k_i^+ = \text{The next highest integer multiple.}$

**Step D.** Determine new estimates of the $k_i$ by evaluating the cost penalty incurred by using $k_i^- T$ and $k_i^+ T$ as the production cycle for product $i$. The cost for each product as a function of $k$, $C_i(k)$, is

$$C_i(k) = \frac{u_i}{k_i T} + 0.5h_i \hat{\alpha}_i d_i k_i T \left( 1 - \frac{\hat{\alpha}_i d_i}{p_i} \right).$$ \hspace{1cm} (16)

The new $k_i$ are chosen by

$$k_i = k_i^- \text{ for } C_i(k_i^-) \leq C_i(k_i^+),$$ \hspace{1cm} (17a)

$$k_i = k_i^+ \text{ for } C_i(k_i^+) \leq C_i(k_i^-).$$ \hspace{1cm} (17b)

The IM procedure is exactly the same as Doll and Whybark's first four steps, except for the fact that the demand rate is replaced by the expected fill rate times the demand rate. Note that Doll and Whybark's approach is aimed at a deterministic situation in which all demand will be filled. The situation modeled in this study is characterized by stochastic demand of which only a certain portion will be filled. We use a part of their procedure to determine the integer multiples. The actual length of the cycle in their procedure is also based on cost minimization, while a more capacity oriented determination of the cycles is appropriate in this case. As discussed before, the traditional ELSP procedures sometimes lead to a shorter cycle time with higher demand levels, whereas a longer cycle time is required.

In Step 5, the heuristic checks whether any of the multiples has changed in the previous step. If so, then the $I_i$'s must be adapted so that the expected fill rates remain the same. In general, the increase in inventory is less than proportional to the increase in $k_i$, since the coefficient of demand variation is smaller over longer time intervals.

In Step 6, the expected profit is calculated and compared to the calculated profit in the previous iteration. Since the function is rather flat around the maximum, the procedure is stopped if the increase in profit is less than $\gamma / \beta$. Determining this criterion is a strategy-related issue of the company. If this ratio is set very large, a lower overall fill rate with less investment in inventory is obtained (high return on investment). If the ratio is set very low, a higher overall fill rate is reached at the expense of an increase in inventory. In this last case, the profit may be hardly different from the profit with a low ratio, but the ROI will be much lower. Note that a considerable increase in target inventory level will be needed to increase the service level marginally, since the tail of the demand distribution function needs to be considered. The target cycle time will not be increased extremely, because the extra quantity to be produced each cycle is only marginal; due to the small increase in fill rate, most of the extra inventory will be left at the end of the cycle.

In Step 7, the product which has the greatest expected contribution per unit of capacity for the next unit stocked (increase of $I_i$) is selected. The target inventory level of the selected product is increased by one unit.

In Step 8, the target inventory level of the product that was selected in the previous step, is increased by 1. After this, the heuristic continues with Step 3, where the cycle times are recalculated with the new value of one of the target inventory levels. When the heuristic is finished, it provides the operations manager with the target inventory levels and the corresponding target cycle times.
5. Experimental tests

As mentioned in Section 2, the literature is replete with scheduling policies which handle more or less different versions of the problem compared to the one addressed in this paper. The existing methods, however, are not capable of coping with very high cumulative demand rates (relative to production rate), in combination with stochastic demand. Of the policies reported in the literature, the variable cycle times based procedure developed by Leachman and Gascon (1988) has been shown to be the most successful one. Our present work extends their policy in two major ways. These extensions enable the policy to perform under very high demand levels. First, our policy not only determines target cycle times (in the top level of the hierarchical model – capacity coordination phase), but also simultaneously selects target service levels. Second, our policy strictly enforces the principle of stable cycle times (at the operational level – bottom level of the hierarchical model), as described earlier. These two extensions are captured in a hierarchical planning system with an embedded heuristic to determine the system parameters based on profit maximization subject to capacity restrictions and minimum service level requirements. Profit maximization is used since lost sales result in reduced contribution to profit. A cost minimization approach would not be able to fully capture this effect.

The research questions rising from the design of our hierarchical procedure refer to these two extensions. These questions are:

(a) Is it better to keep the cycle times stable than to vary them in the short run, and how does this depend upon the aggregate demand level?

(b) Given a policy of stable cycle times, does the approximate model which serves to determine these cycle times adequately capture the real system behavior, and, therefore, is it suitable to determine those cycle times with this model?

We carried out extensive simulation experiments to investigate the performance of our hierarchical procedure with respect to these two research questions. The experiments involved five items produced on a single machine. Demand was generated on a period by period basis and drawn from a normal distribution. The ratio of the cumulative demand and the available capacity varied from about 80% to 105% not counting set-up time. Set-up times were independent of sequence. Two performance measures were used. The first one was the relevant profit, defined as the net contribution of the sold products minus the costs for set-up and inventory. The second performance measure was the fill rate of each of the individual products. Both performance measures can be calculated from the aggregate model (as an estimate) and from the actual simulation. The similarity between the estimate and the actual values served as a performance measure for the quality of the aggregate model. Three sets of tests were performed.

In the first set of tests, a stable cycle times policy (SCT) without capacity coordination was compared to a variable cycle times policy (VCT) based on the heuristic by Leachman and Gascon (1988). The results of these tests show that SCT produces considerably greater profits compared to VCT when the cumulative demand level exceeds 85% of the available capacity. SCT outperforms VCT by 9.1% to 12.1%. As the demand level is decreased, relative to capacity, the performance of VCT improves, which supports the observations by Gascon et al. (1992).

In the second set of tests, the internal validity of the hierarchical model (thus including capacity coordination) was investigated. The results demonstrate that both the relevant profit (as expressed in Eq. (4)) and the distribution of capacity over the various products are well controlled by the hierarchical procedure. This means that the actual capacity allocation in the stochastic, lost sales environment closely resembled the target allocation in the aggregate deterministic model, based on profit considerations. The difference between the estimated relevant profit and the actual relevant profit ranged from 0.2% to 2.2%. This is important from a practical point of view, since this will enable the operations manager to determine targets at a tactical decision level which will be realized if the hierarchical procedure with stable cycle times is used.

In the first two sets of tests, only the demand
level was varied in order to analyze the performance of the various procedures in detail. In order to obtain more general conclusive results, a third set of tests was performed, comparing the hierarchical procedure and the variable cycle times heuristic under many different operating conditions, including variations in demand level, demand variance, demand distribution, set-up times and contribution margins of the various products. The utilization levels (ratio of customer demand and capacity) varied from 91.7% to 100%. The results were analyzed using ANOVA in a randomized block design. Significance, taking profit as the performance measure, was obtained at the 0.005 level. On average, over all 180 observations, the actual profit under the proposed hierarchical procedure was 22% more than the profit realized under the VCT policy.

6. Discussion of the hierarchy

The model presented in this text is a two-tiered hierarchical model. The hierarchical approach used in the development of this model has some specific characteristics. We refer especially to the setting of the control parameters at the upper level and the more myopic view at the operational decision level. In this section we will analyze the hierarchical structure of the model presented in this text. In this analysis, some of the insights provided by Schneeweß (1995) and Bertrand et al. (1990) will be used.

First, let us consider the top level of the model. The top level decides on the control parameters of the system (target inventory level and target cycle time). There is however a difference in the nature of each of these control parameters. The target inventory level decision is a final decision, i.e., it has a direct consequence for the physical production system. The production system always produces up to the target set by the capacity coordination function. The decision on the target cycle time is however not final, since this cycle time will not be realized exactly by the production system. In line with Schneeweß (1995), we call this decision factual, i.e., once the decision is taken by the top level, it is a fact for the bottom level and has to be taken into account in its own decision process.

The top level takes its decision based upon an aggregate model of the bottom level. This aggregate model consists of Eq. (4), (7), (8), and (9). An aggregate representation may be used, if the steady state performance of the detailed model is the same as the steady state performance of the aggregate representation (Chandy et al., 1975). In the design of the hierarchical procedure presented in this text, this consistency has been tested by a series of simulation experiments. In these experiments, performance measures were evaluated for the aggregate and the detailed models (such as profit, individual product service levels and cycle times).

In the aggregate model, some detailed information is not included. First, there is an aggregation concerning the modeling of time. In the aggregate model, time buckets are not distinguished and only rates (for demand and production) are considered, while in the detailed model at the bottom level time periods are distinguished and individual events (order arrivals, setups, etc.) are considered. Secondly – and an extension of the time aggregation – there is an aggregation of the demand process. Demand is modelled as normally distributed demand over the cycle time. This is more or less independent of the actual demand process. To illustrate this, let us suppose the actual demand process follows a Poisson distribution. Since only the demand during the cycle time influences the performance of the system, we may use the characteristic that the Poisson distribution – and a number of other distributions – approximates the normal distribution if it is accumulated over a certain time period. Obviously, this aggregation is only valid if the cycle time is sufficiently long. Thirdly, the interactions between the products are not modeled at the aggregate level. At that level, it is assumed that it will be possible to maintain the individual product cycle times at the detailed level. Since these cycle times are intended to serve as a target for the lower level, this assumption consists of two premises of the bottom level decision procedure. The first premise is that the objective of the bottom level is to minimize the difference be-
tween the actual cycle times and the target cycle times (the objective function of the bottom level). As mentioned in the previous section, the performance of a number of possible decision procedures has been evaluated in order to determine the decision procedure at the bottom level. This evaluation was based upon the difference between the target (cycle times and service levels) determined by the top level and the actual realization. The second premise of the bottom level is that it accepts lost sales, i.e., demand gets lost if it cannot be delivered from the actual inventory. Backordering of demand is not allowed.

Thus, in the aggregate model, the lower level decision and the primary process are represented. It is anticipated what will be the expected reaction of the bottom level. All decisions of the bottom level (sequence and lotsize) take effect in the object system and are — by definition — final (Schneeweiss, 1995).

The consistency exposed here is important in the design of the decision procedures. The aggregate model is based upon a number of assumptions about the bottom level behavior. These assumptions are however in the design process used as premises in the design of the detailed model, leading to consistency in the planning and control hierarchy. The use of these consistency requirements in the design process is in line with the methodology presented by Bertrand et al. (1990).

7. Conclusions

We presented a two-tiered hierarchical approach for scheduling production in multi-item single machine systems, facing very high levels of stochastic demand. Additionally, different contribution margins for each of the products are considered. At the top level, the proposed model embeds a profit maximizing capacity coordination heuristic for determining system parameters (i.e., target cycle times and inventory levels) in the long-term, subject to capacity and service level constraints. The heuristic strictly enforces a stable cycle times policy and determines inventory targets. At the bottom level, the proposed model implements a fixed, order up-to sequencing regime in the short term. Extensive simulation tests comparing the proposed approach to a policy which varies the cycle times indicates the superiority of the stable cycle times policy in terms of total profit and fill rate, at the high demand levels. Results for the deterministic case may be different and have not been discussed in this paper.

The results appear to validate the internal consistency of the proposed approach in that the actual profit and fill rate performance results, from the simulation tests, are extremely close to the estimated values using the model. This result is noteworthy because it enhances the practical usefulness of the model for tactical decision making. Note again that the aggregate model contains estimations of the detailed processes and therefore modelling errors are introduced.

The central aspects of the model, stable cycle times and fixed sequences, are being implemented in a glass container manufacturing company. The initial results are very encouraging. A couple of possible extensions can be suggested. In some cases it may be necessary to distinguish between classes of customers, in terms of sales volume. Thus the two-tiered model can be extended to incorporate a capacity rationing level. The proposed model assumes a perfect market for the products and, hence, may be extended to handle the case where profit margins depend upon the customer, rather than the product. Furthermore, at the operational level it is assumed that a sequence can easily be determined. However, this is not straightforward or even possible in all cases. Research in this area, following this work and the work of Gallego (1990), is encouraged.

Appendix. Use of the partial expectation

The partial expectation function $G(k)$ is defined by Brown (1963) as

$$G(k) = \int_k^{\infty} (t-k) p(t) \, dt$$  \hspace{1cm} (A.1)

where

$p(t) =$ Standard normal density function.
This is graphically illustrated in Fig. A.1. This figure shows the curve of the density function. The partial expectation $G(k)$ is the expectation of the hatched area.

Since the function $G(k)$ is related to the standard normal density function $p(t)$, $\sigma_i T_i^* G(k)$ is the expected quantity short over the cycle time $T_i^*$.

The expected inventory level at the start of a new production run (which is at the end of the cycle time) is

$$\int_{-\infty}^{k} (k-t)p(t) \, dt$$

(A.2)

This is the expected value of the hatched area in Fig. A.2.

Because of the symmetry in the normal distribution density function, expression (A.2) can be written as

$$\int_{-k}^{\infty} (t-k)p(t) \, dt$$

(A.3)

which again is $G(-k)$.

Since $p(k)$ is the standard normal density function ($\mu = 0$ and $\sigma = 1$), the expected quantity short during the cycle time as a function of $I_i^*, T_i^*, d_i$ and $\sigma_i$ is (compare Fig. A.1 and expression (A.1))

$$\sigma_i T_i^* G \left( \frac{I_i^* - d_i T_i^*}{\sigma_i T_i^*} \right).$$

(A.4)

Similarly, the expected inventory at the start of a new production run as a function of the same variables is (compare Fig. A.2 and expressions (A.2) and (A.3))

$$\sigma_i \bar{T}_i^* G \left( \frac{d_i T_i^* - I_i^*}{\sigma_i T_i^*} \right).$$

(A.5)

Acknowledgement

Prof. Ch. Schneeweiss is acknowledged for the discussion which led to Section 6 of this paper.

References


