OR Spektrum (1997) 19: 77–85

OR Spektrum

A capacity-oriented hierarchical approach to single-item and small-batch production planning using project-scheduling methods

Birger Franck, Klaus Neumann, Christoph Schwindt

Institut für Wirtschaftstheorie und Operations Research, Universität Karlsruhe, Kaiserstr. 12, D-76128 Karlsruhe, Germany
(Tel. 0721-608 3809/3808, Fax: 0721-608 3082, www: http://www.wior.uni-karlsruhe.de, e-mail: neumann@wior.uni-karlsruhe.de)

Received: 12 March 1996 / Accepted: 22 July 1996

Abstract. Most production planning and control (PPC) systems used in practice have an essential weakness in that they do not support hierarchical planning with feedback and do not observe resource constraints at all production levels. Also, PPC systems often do not deal with particular types of production, for example, low-volume production. We propose a capacity-oriented hierarchical approach to single-item and small-batch-production planning for make-to-order production. In particular, the planning stages of capacitated master production scheduling, multi-level lot sizing, temporal and capacity planning, and shop floor scheduling are discussed, where the degree of aggregation of products and resources decreases from stage to stage. It turns out that the optimization problems arising at most stages can be modelled as resource-constrained project scheduling problems.

Zusammenfassung. Die meisten in der Praxis eingesetzten Produktionsplanungs- und Steuerungssysteme (PPS-Systeme) besitzen den Nachteil, daß weder eine hierarchische Planung mit Rückkopplungen ermöglicht wird, noch die Ressourcenbeschränkungen auf allen Planungsstufen beachtet werden. Außerdem sind PPS-Systeme meist nicht auf die Anforderungen verschiedener Organisations- und Fertigungstypen, z.B. der Fertigung kleiner Stückzahlen, zugeschnitten. Wir behandeln einen Ansatz für die hierarchische Planung von Einzel- und Kleinserienfertigung bei Kundenauftragsfertigung unter Berücksichtigung beschränkter Ressourcen. Insbesondere werden die Stufen der kapazitierten Hauptproduktionsprogrammplanung, der mehrstufigen Losgrößenplanung, der Termin- und Kapazitätsplanung sowie der Maschinenbelegungsplanung betrachtet, wobei das Niveau der Produkt- und Ressourcenaggregation jeweils von Stufe zu Stufe abnimmt. Die meisten Optimierungsprobleme, die hierbei auf den einzelnen Planungsstufen auftreten, können als ressourcenbeschränkte Projektplanungsprobleme modelliert werden.

Correspondence to: K. Neumann

Key words: Single-item and small-batch production, make-to-order production, hierarchical planning, project scheduling

Schlüsselwörter: Einzel- und Kleinserienfertigung, Kundenauftragsfertigung, Hierarchische Planung, Projektplanung

1. Introduction

The production planning and control systems (*PPC systems*) currently used in practice generally have some disadvantages, which result in large work-in-process inventories, long throughput times, and deadlines frequently being exceeded. PPC systems do not allow for a useful hierarchical planning process with feedback, do not take account of the limited availability of resources at all production levels, and often do not support production environments different from ordinary batch production. As to the latter point, increasing international competition has forced many companies to give more attention to special requests of customers and have led to small batch sizes and a greater variety of products. We shall therefore concentrate on single-item and small-batch production, where we deal with *make-to-order production*, which is typical of single-item and small-batch production. We now review some of the literature pertaining to the field of hierarchical production planning.

After the fundamental work of Hax & Meal (1975), several approaches to hierarchical production planning have been proposed, cf. Dempster et al. (1981), Steven (1994), Carravilla & de Sousa (1995), and Stadtler (1996). Also, Schneeweiß (1989, 1992, 1994, and 1995) has done much pioneering work in that area. To observe scarce resources at all production levels, basic concepts of a hierarchical capacity-oriented PPC system have been devised by Drexl et al. (1994b) and further discussed in Günther & Tempelmeier (1995). The latter approach includes the planning and control stages

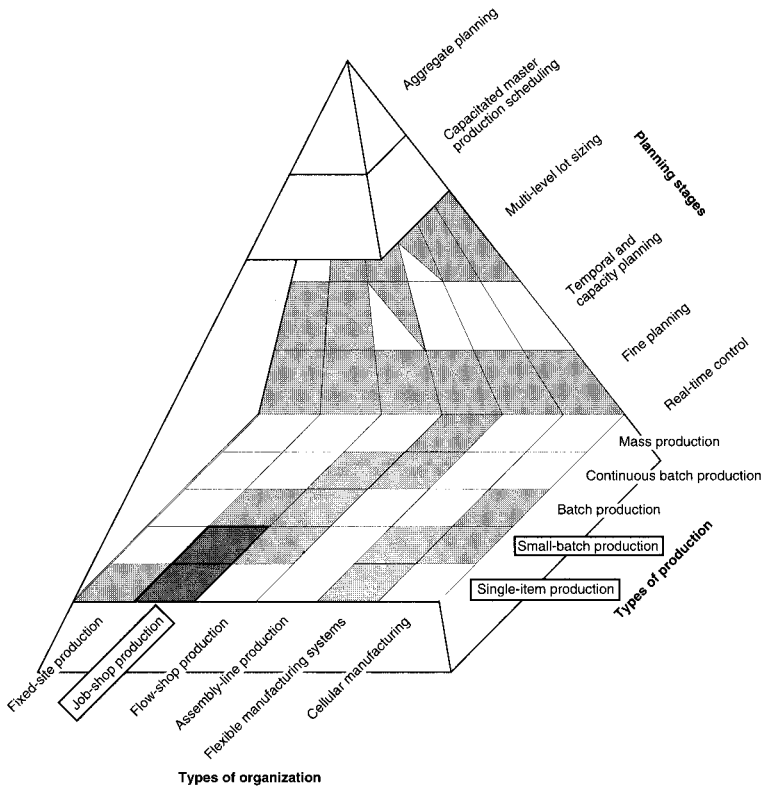

Fig. 1. Pyramid of segment-oriented hierarchical production planning

- Aggregate planning,
- Capacitated master production scheduling,
- Multi-level lot-size and capacity planning,
- Fine planning of individual production segments, and
- Real-time control.

These different planning stages, as well as types of organization and production, are discussed in more detail by Neumann (1996). In this discussion, *production segments* represent combinations of organization and production types, and form the base of a pyramid (see Fig. 1). Possible production segments are illustrated by dark squares of the base of the pyramid in Fig. 1. The stage of lot-size and capacity planning is decomposed into two stages: lot sizing and temporal and capacity planning. Combinations of production segments and planning stages represent cuboid-like parts of the pyramid, whose projections upon the base and a lateral face of the pyramid are depicted in Fig. 1. The darker areas of the lateral face show which planning stages are to be performed for the individual production segments. If a field is not fully but triangularly shaded, the corresponding planning stage is performed for the respective production segment only in some cases. For example, lot sizing is performed for assembly-line production (combined with mass or continuous batch production) in case of an economic lotsizing and scheduling problem but not in automobile assembly.

Single-item production and *small-batch production* form production segments when they are combined with *job-shop production*. We shall present an approach to *capacitated hierarchical planning* for these production segments, which is based on resource-constrained project scheduling and capacitated multi-level lot sizing.

2. Overview of the individual planning stages

We now provide an overview of the individual planning stages mentioned in Sect. 1. These planning stages will be discussed in more detail in the subsequent sections.

Aggregate planning refers to the whole of the enterprise and its production program and is based on long- and medium-term trends. Work force levels have to be matched with the demand forecasts, where a general strategy of a firm is often to keep the work force level low and as constant as possible. This strategy avoids frequent and expensive changes in the size and composition of the work force level at lower planning stages. Groups of final products, instead of single items, are managed over a planning horizon from one to three years at the stage of aggregate planning. However, we shall not discuss this top planning stage because it has little connection with make-to-order production.

Short-term forecasts of future demand for final products and firm customer orders are used to determine a *master production schedule* (*MPS*). Throughout this paper, we deal with master production schedules that are capacitated. Since make-to-order production is typical of low-volume production environment, we shall consider only customer orders and not demand forecasts.

The MPS aims at matching the production program given by firm customer orders with the resources avail-

Table 1. Overview of the planning stages and the degrees of aggregation

Planning stages	Planning horizon/periods	Resources	Items to be scheduled	Output
Aggregate planning	1–3 years/quarters	whole enterprise	demand forecasts for families of final products	medium-term resource requirements, customer orders
Capacitated master production scheduling (MPS)	year/months	work centers, main branches	final products and main components	milestones for final products
Multi-level lot sizing	quarter/weeks	groups of uniform machines	intermediate products	production orders (lots) for intermediate products
Temporal and capacity planning	week/shifts	groups of uniform machines	production orders for individual products	time-phased production orders (jobs)
Fine planning	working day/hours	individual machines or groups of identical machines	time-phased production orders	machine schedules
Real-time control	hour/minutes	individual machines	individual parts	feedback

able. Resources are combined in work centers or main branches of production. The primary requirements for final products are translated into gross requirements for main components (or main products) at lower production levels, exploiting the product structure of the company. The planning horizon is usually about one year comprising twelve periods of one month each. In contrast to make-to-stock production, costs depending on lot sizes are of minor importance in a make-to-order environment. Instead, costs related to the consumption of resources are of greater importance. A constant and low work load is best for ensuring feasible solutions at the subsequent stages, which, in addition, results in a small cost of resources consumption. Therefore, we formulate the production planning problem as a resource-levelling project scheduling problem where delivery dates of customer orders have to be observed. The resulting MPS provides *milestones* for when to produce the customer-ordered final products at the latest and the corresponding resource requirements.

Multi-level lot sizing deals with a general product structure, where final and main products are decomposed into intermediate products. The resources are combined in groups of uniform machines (for example, lathes that may differ in speed), and associated workers. The planning horizon is usually about three months comprising 13 periods or weeks, respectively. The result of this planning stage is the specification of lot sizes for the intermediate products (also called *production orders*), with resource constraints observed.

At the stage of *temporal and capacity planning*, the intermediate products are further decomposed into individual products. For each week (period of lot sizing), completion times for the lots of individual products are calculated and the resources needed for processing the lots are determined. The time elapsed up to the completion of a lot must be specified in terms of a precise number of shifts, i.e. it must be *shift-precise*. This means that *time-phased production orders* (also called *jobs*) are fixed. At this stage, if a feasible schedule cannot be identified (that is, the period of one week is not sufficient to process all jobs on the machines available), we return to the previous stage and determine new lot sizes based on modified resources. This will be discussed later on in more detail.

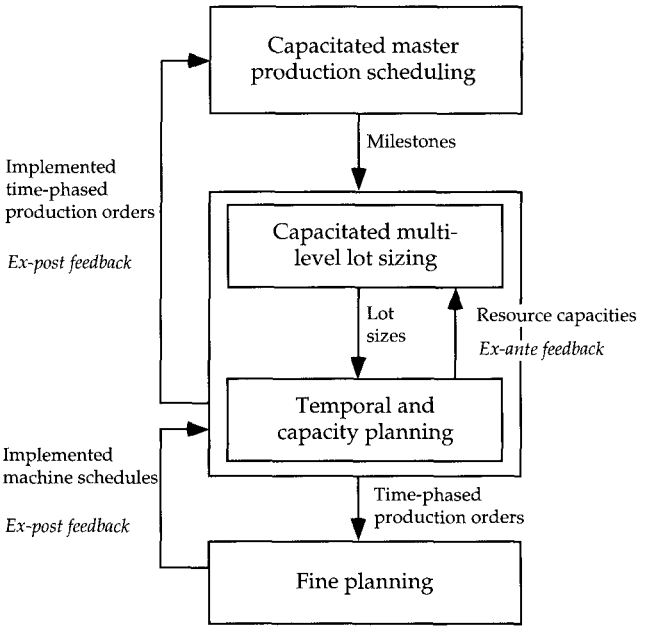

Fig. 2. Hierarchical production planning for make-to-order production

In the case of single-item and small-batch production, the stage of fine planning deals with *shop floor scheduling*, that is, how to process the jobs through the individual machines in a prescribed sequence such that due dates are met. The due date of a job is defined to be the completion time of the corresponding time-phased production order determined at the preceding stage. The planning horizon is usually one working day with the unit of time often being a number of minutes or possibly even about an hour. Shop floor scheduling requires the solution of a job-shop

scheduling problem or a resource-constrained project scheduling problem.

The final *real-time control* monitors and controls the processing of jobs minute after minute where, in practice, an electronic *leitstand* (cf. Drexl et al. 1994 a) is often used. We will not discuss this further in what follows because it is beyond the actual planning stages.

The aggregation of products and resources, as well as the length of planning horizons and periods, may differ from the values proposed above (cf. Konz 1989 and Schneeweiß 1989, 1992). Sometimes, the stage of fine planning is dropped, or performed manually. Table 1 summarizes the time horizons and degrees of aggregation at the individual planning stages.

Since the production planning environment is dynamic, *rolling horizons* should be used. This means that only the first-period results of a planning stage are exploited at the following stage. The full, say, T-period problem is rerun each period to compute new first-period results. When the horizon is moved forward one period, changes generally occur in a schedule. *Nervousness* of planning results may be caused, for example, by updated forecasts, late delivery of primary products, or absence of key personnel. The length of a period should be small enough (compared with the time horizon) to alleviate such nervousness.

Successive planning stages have to be coordinated with each other, where *top-down influence* as well as *bottom-up influence* occur (cf. Schneeweiß 1992, 1995). Top-down influence implies that the results of some stage represent instructions for the following stage. Bottom-up influence means that the results of a stage may cause some modification of the planning process at a previous stage before these results have been implemented (that is, before they have become final decisions), if a performance target at the later stage cannot be met (*ex-ante feedback*). For example, this may happen at the two stages lot sizing and temporal and capacity planning (darker box in Fig. 2), which will be discussed in more detail later on. Another possibility is that the results at a lower stage are employed at an upper stage after these results have been implemented (*ex-post feedback*). An example of the latter type of feedback is the use of rolling horizons. Fig. 2 illustrates hierarchical production-planning for make-to-order production where top-down and bottom-up influence occur.

3. Capacitated master production scheduling

At the MPS stage, we aim at scheduling the production of customer-ordered final products and main components such that the resource requirements of work centers or main branches are as constant (in time) as possible. Each customer order consists of a set of ordered final products and respective order quantities. All products belonging to one and the same customer order have to be delivered at the same prescribed month-precise delivery date. The product structure of the company may be given by bills of materials, a gozinto graph, or product trees. From order quantities and the product structure, the gross requirements of main components can be determined by a bills of materials explosion (cf. Nahmias 1993 and Neumann 1996).

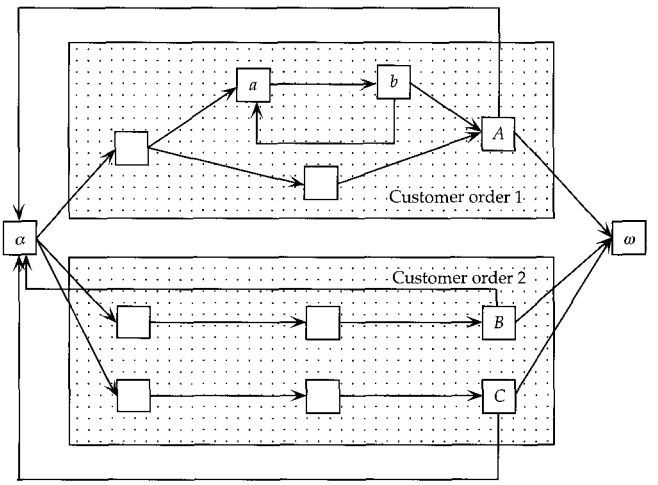

Fig. 3. Multi-project network

In the following, the above production scheduling problem for final products and main components will be modelled as a project scheduling problem, which requires the construction of a *project network*.

In make-to-order production, each customer order can be regarded as a project to be performed. To determine an MPS where resource capacity is observed, we construct a project network for each customer order. These individual project networks are joined together to make a *multi-project network*. Each final product (that belongs to some firm customer order) or main component considered at the present planning stage is viewed as an *activity* of the project. For project planning and scheduling it is recommended to use *activity-on-node networks*, where each activity j is assigned to a node j of the network and the weight b_{ij} of an arc $\langle i, j \rangle$ corresponds to a minimum (or maximum) time lag between the start of activities i and j if b_{ij} is nonnegative (or negative). For the construction of such a network we refer to Neumann (1996) and Neumann & Schwindt (1995).

To manufacture or assemble the gross requirement for a product j, some time D_j is needed and some (renewable) resources are required. The execution time or duration D_j of the corresponding activity j results from summing up the respective processing times of product j itself and of the components of product j at lower levels of the product structure, where a surcharge for transportation and handling may be added. The resources required are determined by summing up the respective machine units and workers needed. To avoid peak demand for resources, the resource requirements for product j are assumed to be distributed uniformly over the execution time D_j so that the resource demand rates are constant. Note that the assumption of constant resource requirements may lead to an underestimation of the consumption of resources in some periods. That drawback can be offset by adding a surcharge to the constant resource requirements or by linking the stages of master production scheduling and lot sizing by an ex-ante feedback approach. The latter represents an area of future research and is not discussed in this paper. A more detailed distribution of resources over time will be considered at

the later planning stage of temporal and capacity planning. We now discuss a multi-project network model of these ideas.

The multi-project network contains an initial node α and a terminal node ω (connected with the sources or sinks, respectively, of the individual project networks). A delivery date or deadline δ_j for some product j can be modelled by a maximum time lag of size $\delta_j - D_j$ between the dummy activity corresponding to initial node α and the start of activity j. The maximum project duration T prescribed is the maximum of the deadlines of all activities. A temporal analysis for the multi-project network provides earliest and latest start times ES_j and LS_j, respectively, as well as earliest and latest finish times EF_j and LF_j, respectively, for all activities j (cf. Neumann & Morlock 1993).

Figure 3 shows a simple multi-project network, which models the manufacture and assembly of three final products, comprising two individual project networks each corresponding to a customer order. The activities or nodes A, B, and C represent final products. The backward arc from node b to node a corresponds to a maximum time lag between the start of activities a and b. The backward arcs from A to α, from B to α, and from C to α mean that there are prescribed delivery dates for the final products A, B, and C.

Suppose that the multi-project network consists of n activities or nodes, respectively, 1, ..., n and let the fictitious activities 0 and $n+1$ correspond to initial node α and terminal node ω, respectively. Let $\mathcal{J}=\{0, 1, \ldots, n, n+1\}$ be the set of activities and let $\mathcal{P}_j$ be the set of the (immediate) predecessors of activity j. Moreover, let $\kappa=1, \ldots, K$ denote the (renewable) resources, and let R_κ be the amount of resource κ available and $r_{j\kappa}$ be the amount or resource κ required for the processing of activity j. Machines, workers, and tools can be modelled as (renewable) resources. We introduce the binary variables

$$x_{jt} := \begin{cases} 1, & \text{if activity } j \text{ is completed at the end of period } t \\ 0, & \text{otherwise} \end{cases} \qquad (j \in \mathcal{J};\ t = 0, \ldots, T),$$

where the beginning of the project is said to fall into period 0.

The problem of determining an MPS, which attempts to match the production program (given by customer orders) with the resource capacity available, can be formulated as a *resource-levelling problem* for the multi-project network as follows:

$$\text{Minimize } F(x_{jt} \mid j \in \mathcal{J}, t \in \Delta_j) \tag{1}$$

subject to

$$\sum_{t \in \Delta_j} x_{jt} = 1 \qquad (j \in \mathcal{J}) \tag{2}$$

$$\sum_{t \in \Delta_i} x_{it} \cdot (t - D_i + b_{ij}) \le \sum_{t \in \Delta_j} x_{jt} \cdot (t - D_j) \qquad (j \in \mathcal{J};\ i \in \mathcal{P}_j) \tag{3}$$

$$r_\kappa(t) \le R_\kappa \qquad (\kappa = 1, \ldots, K;\ t = 0, \ldots, T) \tag{4}$$

$$\sum_{t \in \Delta_{n+1}} x_{n+1,t} \cdot t \le T \tag{5}$$

$$x_{jt} \in \{0, 1\} \qquad (j \in \mathcal{J};\ t \in \Delta_j) \tag{6}$$

where

$$\Delta_j := \{EF_j, EF_j + 1, \ldots, LF_j\} \qquad (j \in \mathcal{J}),$$

and

$$r_\kappa(t) := \sum_{j \in \mathcal{J}} r_{j\kappa} \sum_{\tau \in \Delta_j \cap [t, t+D_j)} x_{j\tau} \quad (\kappa = 1, \ldots, K;\ t = 0, \ldots, T)$$

is the amount of resource κ required in period t. Possible objective functions F for the resource-levelling problem are, for example,

$$\max_{t=0,\ldots,T} \max_{\kappa=1,\ldots,K} g_\kappa r_\kappa(t), \quad \max_{t=0,\ldots,T} \sum_{\kappa=1}^{K} g_\kappa r_\kappa(t),$$

$$\sum_{t=0}^{T} \sum_{\kappa=1}^{K} g_\kappa [r_\kappa(t)]^2, \quad \text{and} \quad \sum_{t=0}^{T} \sum_{\kappa=1}^{K} \left| \bar{R}_\kappa - r_\kappa(t) \right|$$

where $g_\kappa \ge 0$ is a weighting factor (e.g. the cost per unit of resource κ) and $\bar{R}_\kappa$ represents some target value for the consumption of resource κ. We now explain the above constraints.

Equations (2) guarantee that activity j is carried out without interruption. Inequalities (3) ensure that the minimum and maximum time lags are observed, and inequalities (4) represent the resource constraints. Inequality (5) guarantees that the maximum project duration T is not exceeded.

Instead of a resource-levelling problem, a so-called *resource-investment problem* may sometimes be more expedient, for example, if some expensive resources are leased, or the company in question wants to outsource some complex intermediate products. Then, the objective function to be minimized is of the form

$$\sum_{\kappa=1}^{K} c_\kappa(R_\kappa)$$

where the resource capacity R_κ is considered a variable and $c_\kappa(\bullet)$ is a nondecreasing cost function (cf. Demeulemeester 1995 and Möhring 1984). We now review some of the recent work reported on resource-levelling and resource-investment problems.

Heuristic procedures for the resource-levelling problem were proposed by Brinkmann & Neumann (1996). For the case of no maximum time lags, heuristics were devised by Harris (1990), Leachman (1983), and Neumann & Morlock (1993), whereas Bandelloni et al. (1994) proposed a dynamic programming approach. All algorithms reported so far do not consider the resource constraints (4). An exact branch-and-bound-based method for the resource-investment problem without maximal time lags was devised by Demeulemeester (1995).

In general, the execution time D_j of an activity j is considerably larger than the sum of the processing times of product j, due to waiting times, which are known only after shop floor scheduling has taken place. Hence, in practice, D_j is found by adding a surcharge (of often up to 500%) to the sum of the processing times. This approach, however, does not account for the dependency of execution times on the utilization of resources. In fact, practical experience shows that the execution times increase heavily with growing utilization (Karmarkar 1987). Schnee-

weiß & Söhner (1995) have used queueing models to determine expected execution times as a function of resource utilization. The latter approach can be employed for estimating a surcharge (depending on resource utilization) to be added to the sum of the processing times.

In the resource-levelling problem, resource utilization can be calculated from the given resource requirements $r_{j\kappa}$ and resource availabilities R_κ. For the resource investment problem, the utilization can be determined analogously prior to each step of the iterative algorithm of Demeulemeester, which consists of the repetitive solution of so-called resource-constrained project scheduling problems with fixed resource availability R_κ. For details we refer to Demeulemeester (1995).

We shall now continue our discussion of individual planning stages by going on to deal with lot sizing.

4. Lot sizing

The MPS provides month-precise milestones for production orders for final products such that the utilization of work centers over time is well-balanced. The following planning stage of lot sizing determines lot sizes for final products and for intermediate products which are capital-, time-, or wage-intensive such that the MPS milestones are observed, the (aggregated) capacities of groups of uniform machines in the work centers are not exceeded, and the sum of setup and (inventory) holding costs is minimized. To find the gross requirements for all products considered at the lot-sizing stage, the multi-level product structure of the company has to be exploited (bills of materials explosion), cf. Nahmias (1993) and Neumann (1996). The planning horizon usually amounts to three months, comprising 13 periods. The lot-sizing stage provides week-precise production orders for intermediate products.

The problem just described can be modelled as a *multi-item, multi-level capacitated lot-sizing problem* and solved approximately by heuristics proposed by Tempelmeier & Derstroff (1993) and Tempelmeier & Helber (1994), cf. also Derstroff (1995) and Helber (1994). The drawback of the underlying lot-sizing models, however, is that the lead times for the products are supposed to be fixed externally, independent of the lot sizes. Moreover, waiting times that are caused by limited resources and will be determined at the stage of temporal and capacity planning are not included in the lead times.

To overcome this disadvantage, we solve iteratively the lot-sizing problem (LS problem) and the subsequent temporal and capacity planning problem (TCP problem), which will be discussed in more detail in Sect. 5. At the beginning, the lead times are assumed to be zero. In each of the following iterations, the lead times for the LS problem are set equal to the production lead times found in the preceding iteration of the TCP problem.

In each iteration, from the solution to the LS problem, lot sizes for the remaining (individual) products manufactured by the company can be obtained by means of the product structure. Lot sizes for the purchased components can be determined by exploiting some appropriate (uncapacitated) inventory model (cf. Neumann 1996).

We now discuss the temporal aspects of capacity planning.

5. Temporal and capacity planning

For each week (that is, each period of lot sizing), temporal and capacity planning provides a shift-precise timing of the purchase or production orders, respectively, for all components, subassemblies, and final products, i.e. all individual products. To carry out the production orders, resources are needed, which represent groups of uniform machines. All production orders (or *jobs*) have to be executed within one week. Thus, we seek to minimize the makespan, that is, the maximum completion time of all jobs.

At the TCP stage, the production orders (jobs) are decomposed into operations, where *operation* O_{ij} corresponds to the processing of job j on an average (individual) machine of group or resource M_i, respectively, and the setup of that machine. The sequence in which the operations of a job have to be carried out (machine sequence for that job) is supposed to be given by process plans. Analogous to the MPS stage (see Sect. 3), the execution of the production orders within one period of lot sizing (one week) can be modelled by a *multi-project network*, where the operations correspond to the nodes, and the weight of an arc with initial node O_{ij} and final node O_{kl} corresponds to the minimum time lag between the start of operations O_{ij} and O_{kl}. The processing time of operation O_{ij} must be set equal to the average setup plus processing time of product j on any (individual) machine of resource M_i. The construction of such a multi-project network, where overlapping operations are permitted and maximum time lags may occur, is discussed by Neumann & Schwindt (1995). In the case of a general (acyclic) product structure, *common parts* (products that are components of more than one other product at a higher level) may occur, which results in hard sequencing problems (cf. Günther 1992 and Neumann & Schwindt 1995).

The TCP problem can be formulated as a *resource-constrained project scheduling problem* where the project duration (makespan) $\sum_{t \in \Delta_{n+1}} t \cdot x_{n+1,t}$ is to be minimized. The constraints of this zero-one programming problem coincide with the constraints (2) to (6) from Sect. 3. Exact methods for solving small problems of that type have been devised by Demeulemeester & Herroelen (1992) and Sprecher (1994). Heuristic procedures were constructed by Kolisch (1995) for the case where maximum time lags do not exist, and by Neumann & Zhan (1995), Brinkmann & Neumann (1996), and Franck (1996) if, in addition to minimum time lags, maximum time lags have to be observed. The resulting schedule also yields the lead times of all individual products manufactured.

Two methods for the integrated solution of the LS and TCP problems in a job shop environment are known from literature: the algorithm of Dauzère-Pérez & Lasserre (1994), cf. also Lasserre (1992), and the algorithm of Lambrecht & Vanderveken (1979). Both approaches are based on a two-stage model where lot sizing and job shop scheduling problems are solved alternately. In the algorithm of

Dauzère-Pérez & Lasserre, the LS problems are solved for fixed job sequences on the machines which have been determined in the previous iteration of the job shop algorithm. Hence, sequence-dependent waiting times on the machines can already be considered at the LS stage. This approach, however, is based on the specific property of the job shop model that any resource (i.e. each machine) can process at most one job at the same time. In this case, a feasible schedule can always be derived from the job sequences, which can be done independently of the processing times (that result from the lot sizes provided by the LS stage). Since, at the present TCP stage, we consider groups of uniform machines as resources, a resource is generally able to process more than one job at the same time. This is the reason why the approach of Dauzère-Pérez & Lasserre cannot be adapted to the LS and TCP problems in question. In the following, we develop an iterative solution procedure for the LS and TCP stages which is based on the methodology of Lambrecht & Vanderveken.

Suppose that, in some week, a feasible schedule cannot be found at the TCP stage. That is, the lot sizes from the LS stage cannot be produced on schedule in this week due to the capacity constraints. Then, the resource capacities are reduced appropriately resulting in a reduction of some lot sizes. The solution procedure requires alternating between the LS and the TCP stages until a feasible schedule is found at the TCP stage and the production lead times computed in two successive iterations are essentially the same.

In more detail, we first determine the start time s and finish time f for each resource M_i, i.e. the minimum start time and maximum completion time, respectively, of any job on an average (individual) machine of resource M_i (note that the period begins at time zero and ends at time one [in weeks]). If for the finish time $f>1$ [weeks], then the capacity of the resource is reduced by $(f-1)/(f-s)$. If that capacity reduction does not lead to a reduction of the lot size of at least one product processed on the resource in question, the capacity is further reduced to that capacity which is required for producing the lots of the current week (determined at the LS stage) minus ε, where ε is a positive constant. Then the lot size of at least one of the latter products decreases by at least one. In principle, this corresponds to the method of Lambrecht & Vanderveken (1979) for production scheduling and sequencing of products with linear product structure.

The next stage is fine planning, consisting in shop floor scheduling of single-item and small-batch production, which we now discuss.

6. Shop floor scheduling

The stage of shop floor scheduling deals with processing the jobs (time-phased production orders) on the individual machines. The planning horizon is one working day and the unit of time (period length) is one hour or several minutes. The shift-precise completion times of the production orders from the TCP stage are used as due dates d_j of the respective jobs j at the stage of shop floor scheduling. Let C_j be the completion time of job j at the present stage. Then $\alpha_j(d_j-C_j)$ is an earliness cost when job j is completed early and $\beta_j(C_j-d_j)$ is a tardiness cost when job j is completed late, where $\alpha_j \geq 0$ and $\beta_j \geq 0$ are the earliness and tardiness cost, respectively, per unit of time. Moreover,

$$h_j(C_j) := \begin{cases} \alpha_j(d_j - C_j), & \text{if } C_j \leq d_j \\ \beta_j(C_j - d_j), & \text{if } C_j > d_j \end{cases}$$

represents a penalty cost for job j. Possible objective functions to be minimized are

(a) $\max_j h_j(C_j)$ (maximum penalty cost)

(b) $\sum_j h_j(C_j)$ (sum of penalty costs)

(c) $\max_j C_j$ (makespan)

The problem to be solved comprises two parts. First, each operation has to be assigned to a machine which is suitable for processing that operation. In general, there are several suitable machines, differing in speed. These different individual machines are regarded as *modes* in job-shop and project scheduling, and the problem to be solved is called a *mode-assignment problem* (see Kolisch 1995) where the individual machines within one group of uniform machines are to be utilized evenly. Second, an optimal schedule has to be found that minimizes one of the above objective functions. The latter problem can be formulated as a *job-shop problem* (cf. Brucker 1995 or Pinedo 1995) or a *resource-constrained project scheduling problem* (see Neumann & Schwindt 1995).

A heuristic method for solving the project scheduling problem with a penalty-cost objective function has been devised by Serafini and Speranza (1994), which uses a three-step decomposition approach. First, a mode is assigned to each operation. After that, a sequencing and a scheduling problem are solved. The three steps are linked by ex-ante and ex-post feedback, which identify and deal with so-called critical operations (for details we refer to Serafini and Speranza 1994).

A similar hierarchical approach can be used for the multi-mode job-shop problem. Kolisch (1995) has proposed heuristics for the solution of a special mode-assignment problem. For particular objective functions, the jop-shop problem can be solved by the Giffler-Thompson heuristic (compare Neumann 1996 and Schwindt 1997). The priority rules Shortest Slack Time (SST) and Shortest Relative Slack Time (SRST) have turned out to be appropriate for lateness objective functions (that is, objective functions (a) and (b) with $\alpha_j=0$ and $\beta_j=1$). For the minimization of makespan (objective function (c)), the First-In-First-Out (FIFO) rule and the Most Work Remaining (MWR) rule have provided good results.

In contrast to the lateness and the makespan objective functions, the penalty-cost functions (a) and (b) represent, in general, nonregular functions. That is, they are not non-decreasing in the job completion times C_j. Algorithms for the solution of job-shop problems with nonregular objective functions have not yet been proposed.

For the special case where the processing time of an operation does not depend on the specific machine selected (the case of multi-purpose machines), a heuristic that solves a mode-assignment and a job-shop problem simultaneously has been proposed by Hurink et al. (1994). Boctor (1994) discusses how to solve mode-assignment and resource-constrained project scheduling problems simultaneously for the makespan objective function provided that there are only minimum time lags between activities.

7. Conclusions

We have presented a capacity-oriented hierarchical approach to make-to-order production. Most of the optimization problems arising at the planning stages of capacitated master production scheduling, lot sizing, temporal and capacity planning, and shop floor scheduling can be formulated as different types of resource-constrained project scheduling problems with minimum and maximum time lags. For the stages of lot sizing and temporal and capacity planning, a solution procedure which alternates between these two stages turns out to be expedient. For most of the optimization problems considered, efficient heuristic procedures have been proposed recently, as we have mentioned.

We suggest that a fruitful area of further research is likely to be the development of heuristics for the (as yet unsolvable) problems arising from the hierarchical approach: the resource-levelling problem with resource constraints, the resource-investment problem with maximum time lags, the mode-assignment problem where uniform machines are to be utilized evenly, and the earliness/tardiness job-shop problem. Moreover, the planning stages of capacitated master production scheduling and lot sizing should be linked by an ex-ante feedback approach, taking into account the effect of setup times (resulting from lot sizing) and waiting times (determined by temporal and capacity planning) on the execution times of the products used at the MPS stage as well as taking into account the consequences of stipulating constant resource requirements at the MPS stage. Also, similar hierarchical approaches to different types of production should be developed. The authors plan to report on some of these approaches elsewhere.

Acknowledgement. Sincere thanks are due to the anonymous referees. Their comments led to many improvements in this paper.

References

1. Bandelloni M, Tucci M, Rinaldi R (1994) Optimal resource leveling using non-serial dynamic programming. EJOR 78:162–177
2. Boctor FF (1994) Heuristics for solving projects with resource restrictions and several resource-duration modes. Intern J Prod Res 31:2547–2558
3. Brinkmann K, Neumann K (1996) Heuristic procedures for resource-constrained project scheduling with minimal and maximal time lags: the resource-levelling and the minimum project-duration problems. J Decision Systems 5:129–155
4. Brucker P (1995) Scheduling Algorithms. Springer, Berlin
5. Carravilla MA, de Sousa JP (1995) Hierarchical production planning in make-to-order company: A case study. EJOR 86:43–56
6. Dauzère-Pérez S, Lasserre J-B (1994) Integration of lot sizing and scheduling decisions in a job-shop. EJOR 75:413–426
7. Demeulemeester EL (1995) Minimizing resource availability costs in time-limited project networks. Mgmt Sci 41:1590–1598
8. Demeulemeester EL, Herroelen WS (1992) A branch-and-bound procedure for the multiple resource-constrained project scheduling problem. Mgmt Sci 33:1803–1818
9. Dempster MAH, Fisher ML, Jansen L, Lageweg BJ, Lenstra JK, Rinnooy Kan AHG (1981) Analytical evaluation of hierarchical planning systems. Oper Res 29:707–716
10. Derstroff M (1995) Mehrstufige Losgrößenplanung mit Kapazitätsbeschränkungen, Physica, Heidelberg
11. Drexl A, Eversheim W, Grempe R, Esser H (1994a) CIM im Werkzeugmaschinenbau – Der Prisma-Montageleitstand. ZfbF 46:279–295
12. Drexl A, Fleischmann B, Günther H-O, Stadtler H, Tempelmeier H (1994b) Konzeptionelle Grundlagen kapazitätsorientierter PPS-Systeme. ZfbF 46:1022–1045
13. Franck B (1996) Empirische Untersuchungen von Prioritätsregeln für die ressourcenbeschränkte Projektplanung mit zeitlichen Minimal- and Maximalabständen. In: Kleinschmidt P, Bachem A, Derigs U, Fischer D, Leopold-Wildburger U, Möhring R (eds) Operations Research Proceedings 1995. Springer, Berlin, pp 144–149
14. Günther H-O (1992) Netzplanorientierte Auftragsterminierung bei offener Fertigung. ORS 14:229–240
15. Günther H-O, Tempelmeier H (1995) Produktion und Logistik. Springer, Berlin
16. Harris RB (1990) Packing method for resource leveling (Pack). J Constr Eng Mgmt 116:39–43
17. Hax AC, Meal HC (1975) Hierarchical integration of production planning and scheduling. In: Geisler M (ed) Logistics, TIMS Studies in Management Science, Vol. 1. Elsevier, Amsterdam, pp 53–69
18. Helber S (1994) Kapazitätsorientierte Losgrößenplanung in PPS-Systemen. Poeschel, Stuttgart
19. Hurink J, Jurisch B, Thole M (1994) Tabu search for the job shop scheduling problem with multi-purpose machines. ORS 15:205–215
20. Karmarkar US (1987) Lot sizes, lead times and in-process inventories. Mgmt Sci 33:409–418
21. Kolisch R (1995) Project Scheduling under Resource Constraints. Physica, Heidelberg
22. Konz H-J (1989) Steuerung der Standplatzmontage komplexer Produkte. Doctoral Dissertation, RWTH Aachen
23. Lambrecht MR, Vanderveken H (1979) Production scheduling and sequencing for multi-stage production systems. ORS 1:103–114
24. Lasserre J-B (1992) An integrated model for job-shop planning and scheduling. Mgmt Sci 38:1201–1211
25. Leachman RC (1983) Multiple resource leveling in construction systems through variation of activity intensities. Nav Res Log Quart 30:187–198
26. Möhring RH (1984) Minimizing costs of resource requirements in project networks subject to a fixed completion time. Oper Res 32:89–120
27. Nahmias S (1993) Production and Operations Analysis. Irwin, Homewood
28. Neumann K (1996) Produktions- und Operations-Management. Springer, Berlin
29. Neumann K, Morlock M (1993) Operations Research. Carl Hanser, München
30. Neumann K, Schwindt C (1995) Projects with minimal and maximal time lags: construction of activity-on-node networks and applications. WIOR Report-447, Institut für Wirtschaftstheorie und Operations Research, Universität Karlsruhe
31. Neumann K, Zhan J (1995) Heuristics for the minimum project duration problem with minimal and maximal time lags under fixed resource contraints. J Intell Manufact 6:145–154

32. Pinedo M (1995) Scheduling. Prentice Hall, Englewood Cliffs
33. Schneeweiß C (1989) Einführung in die Produktionswirtschaft. Springer, Berlin
34. Schneeweiß C (1992) Planung 2 – Konzepte der Prozeß- und Modellgestaltung. Springer, Berlin
35. Schneeweiß C (1994) Elemente einer Theorie hierarchischer Planung. ORS 16:161–168
36. Schneeweiß C (1995) Hierarchical structures in organisations: A conceptual framework. EJOR 86:4–31
37. Schneeweiß C, Söhner V (1995) Ein hierarchisch integriertes Konzept zur Produktionsplanung und -steuerung. Discussion Paper No. 54, Lehrstuhl für Betriebswirtschaftslehre und Unternehmensforschung, Universität Mannheim
38. Schwindt C (1997) A simulated-based comparison of the shifting bottleneck procedure with the algorithm of Giffler and Thompson for the job shop problem. Submitted to ORS
39. Serafini P, Speranza MG (1994) A decomposition approach for a resource constrained scheduling problem. EJOR 75:112–135
40. Sprecher A (1994) Resource-Constrained Project Scheduling. Lecture Notes in Economics and Mathematical Systems Vol. 409. Springer, Berlin
41. Stadtler H (1996) Hierarchische Produktionsplanung. In: Kern W, Schröder H-H, Weber J (eds) Handwörterbuch der Produktionswirtschaft. 2. Auflage. Schäffer-Poeschel, Stuttgart
42. Steven M (1994) Hierarchische Produktionsplanung. Physica, Heidelberg
43. Tempelmeier H, Derstroff M (1993) Mehrstufige Mehrprodukt-Losgrößenplanung bei beschränkten Ressourcen und genereller Erzeugnisstruktur. ORS 15:63–73
44. Tempelmeier H, Helber S (1994) A heuristic for dynamic multi-term multi-level capacitated lotsizing for general product structures. EJOR 75:296–311