Intl. Trans. in Op. Res. 9 (2002) 779–793

INTERNATIONAL
TRANSACTIONS
IN OPERATIONAL
RESEARCH

On the importance of sequencing decisions in production planning and scheduling

Stéphane Dauzère-Pérès and *Jean-Bernard Lasserre

IRCCyN/Ecole des Mines de Nantes, 4 rue Alfred Kastler, La Chantrerie, BP 20722, 44307 Nantes Cedex 03, France and
**LAAS-CNRS, 7, avenue du Colonel Roche, F 31077 Toulouse Cedex 4, France*
Email: Stephane.Dauzere-Peres@emn.fr [Dauzère-Pérès] and lasserre@laas.fr [Lasserre]

Received 12 November 2001; received in revised form 24 April 2002; accepted 30 May 2002

Abstract

We discuss the traditional hierarchical approach to production planning and scheduling, emphasizing the fact that scheduling constraints are often either ignored or considered in a very crude way. In particular, we underline that the way scheduling is carried out is crucial for the capacity constraints on the lot sizes. Usual methods to handle capacity in theory or in practice are reviewed. Finally, we present an approach that tries to overcome these drawbacks by capturing the shop-floor capacity through scheduling considerations.

Keywords: Production planning, scheduling, sequencing

1. Introduction

The traditional hierarchical approach in production management has long been recognized and accepted in practice. It consists of multiple decision levels (usually three: strategic, tactical and operational) with different characteristics. In particular:

- the higher in the hierarchy the more strategic are the decisions;
- the higher in the hierarchy the more aggregate are the models and the longer the time horizon;
- the decision at some level becomes a constraint or an objective to be satisfied at lower levels;
- each decision level has its own decision models and solving procedures.

These considerations can be found in most textbooks on production management (cf. e.g. Thomas and McClain, 1993 and the references therein). In addition, these hierarchical decision levels often coincide with specialized decision-makers in the companies, which makes this framework even more appealing. This is accentuated by the fact that different departments in a company often have difficulty sharing information, for structural (or more often personal) reasons. This latter phenomenon tends to disappear with the arrival of ERP (enterprise resource planning) systems such as SAP, BAAN or others.

Published by Blackwell Publishers Ltd.

These systems force information sharing because, by definition, information is available from one single database. We found a typical example of such a hierarchical decision process in a factory of a big French mobile-phone company. The marketing department in charge of collecting customer and forecast demands establishes a master production schedule of the finished products after discussion (weekly meetings) with the material management department. This department, using an MRP approach, manages the input and output inventories of the factory, sends purchase orders to the suppliers, and determines the production orders. These orders are then tentatively sequenced in the various shops by the scheduling department whose output is sent to every shop manager, who in turn tries to follow the proposed schedule as closely as possible.

A rationale behind this hierarchical approach is to simplify the overall decision process. In particular, the decision-making procedure of some level does not have to consider 'details' that are unnecessary at this level and, furthermore, only the relevant decision variables are required. Of course, for the overall decision process to be coherent, the decisions taken at some level must make sense. When transmitted to the lower levels as constraints to satisfy or objective to attain, one must be able to provide subsequent 'good' decisions, i.e. compatible with the (higher level) decisions already taken. The compatibility is measured (or evaluated) via some criterion, and some emergency procedures may exist in case this compatibility is not satisfied. When an inconsistency has been detected at some decision level, some higher level decisions have to be re-evaluated according to the new conditions on the system at that time. If those emergency procedures are activated too frequently, it is a sign that the decision process is not coherent.

In hierarchical production planning (HPP), the consistency between different decision levels has been investigated by several authors and the reader is referred to Bitran and Tirupati (1993) for a detailed survey. One may note that, in this case, the two (aggregate and detailed) decision levels are of the same type, i.e. both consider *flows* of products in the workshop.

In this paper, we consider the usual hierarchical approach for planning and scheduling, i.e., we consider the (two) planning and scheduling decision levels. In contrast to HPP, these two decision levels are very different in nature. On the one hand and as in HPP, the planning (or lot sizing) level determines *flows* of products. On the other hand, the scheduling level determines *sequences* of products on the machines. In the planning level, typical models involve continuous variables whereas, in the scheduling level, the models include discrete variables and are primarily combinatorial in nature.

We want to show that, in many cases, the standard hierarchical procedure described above, is not coherent and we underline some important reasons why this is so. This lack of consistency between planning and scheduling decisions has been recognized for some time. For instance, to cite a few authors:

- in Lenstra and Rinnoy Kan (1977) 'There appear to be good opportunities for research on the interface between scheduling and inventory theory. Both . . . have been developed in complete mutual isolation';
- in Smith (1978), 'the lack of appropriate support for managers to produce good master schedule is a major weakness of MRP, and probably the biggest source of disappointment in the performance of such systems';
- in Baker (1993), 'Nevertheless, there are several key weaknesses in the basic MRP framework. These relate to *lot sizes, capacity, planned lead times*, and *uncertainty* . . . It would be desirable to recognize capacity constraints while building the MPS . . . Planned lead times are treated as given . . . It would be desirable to treat lead times as dependent on product mix, shop load, and capacity: in short as

dynamic ... They should be viewed not as inputs to a scheduling procedure but rather as part of the output';

- in Uzsoy, Chung-Yee Lee and Martin-Vega (1994) 'A salient point emerging from this review has been problem areas have been compartmentalized, resulting in inter-related problems being considered in isolation ... A key factor in linking production planning and shop-floor control decisions is the development of accurate methods of modelling manufacturing capacity'.

Moreover, it is recognized that 'planning and scheduling are the two most essential modules of the supply chain' (Seyed, 1996, 1998). Hence, a lot of attention has been devoted to optimizing problems at the two decision levels, but not enough to the interactions between them.

The most common approach used in production planning remains MRP (material requirements planning, see for instance Vollmann, Whybark, and Berry (1997)). The most well-known limitation of this approach is that it works with *infinite capacity.* Moreover, it is unrealistic to use *lead times* (time required to complete a lot of a given item) that do not depend on the size of the lots and on how the lots are processed on the shop floor. Obviously, the size of a given lot influences the time the lot will spend on the machines. As discussed in Karmarkar (1987, 1993), and this is often ignored, the overall amounts of different items to produce also have a direct impact on the lead times of the products. In the next paragraph, we will discuss the influence of shop scheduling on lead times. However, it should be noted that stating that MRP considers infinite capacity is a bit too naïve. In practice, capacity is somehow incorporated in the lead time, since it is well known that often more than 80% (although this figure decreases with just-in-time approaches) of the lead time corresponds to idle time of jobs waiting to be processed or transferred. This idle time illustrates the fact that there is only *finite* capacity. Even though capacity is taken into account in an aggregate way in MRP II (manufacturing resource planning) approaches, it is far from enough to ensure consistency between planning and scheduling decisions (see example in Section 3 below).

For illustration purposes, we briefly sketch in a simple example how scheduling influences the time a lot requires to be completed. If a simple scheduling rule such as SPT (shortest processing time) is used on the shop floor, i.e. when several lots can be processed on a machine, priority is given to the lot which requires the least processing time on the machine. For a given set of lots sent to the shop floor at a given time, the product associated with the lot with the lowest required processing times on the machines will probably be processed before the others, and thus faster. Its lead time will then be rather small. On the other hand, if the size of the lot of this product increases significantly, while the others remain constant, its lead time can become very large. Note that the LPT (longest processing time) priority rule would give the opposite result. In an MRP approach, the same lead time is used, independently of the production quantities and the scheduling policy.

We claim that, in many production contexts, the scheduling decision level should not be considered as a 'lower' (or slave) level in comparison with the planning (or lot-sizing) decision level. This is particularly true when the production is by *lots* (a large class of production environments). The quality of the scheduling policy is as much part of production capacity as the speed or available times of the resources. It is well known that efficient scheduling heuristics can lead to a reduction of more than 20% in the job completion times. For instance, when manual scheduling is performed, expert schedulers will often much better be able to meet due dates than inexperienced ones. This should be taken into account in some way when computing the production plan.

The 'naïve' capacity constraints in most lot-sizing models make sense when the individual items are treated in 'isolation', i.e. when the 'transfer lot' is a single item. Indeed, in this case, compared to the length of the planning horizon, the duration of an elementary operation on an item is negligible, and the transfer time to the next machine is almost 'instantaneous'. Roughly speaking, the lot-sizing model is a 'fluid' model approximation, consistent with the flow of items through the workshop. The total workload on the machines is then a good variable to consider for the capacity restrictions. When the production is by lots, the duration of an elementary operation on a *lot* may not be negligible any more, depending on the size of the lot. The 'fluid' approximation is no longer valid, and the total workload on the machines is not the only relevant variable to consider for capacity restrictions. The sequencing of lots on the machines comes into play and should not be ignored at this decision level. Here, one should distinguish between *sequencing* (determining an ordering of operations on the resources) and *scheduling* (determining a sequence and start times of operations on the resources). Indeed, in view of the (even minor) disturbances that will occur, determining, at the beginning of the horizon, the exact start times (a complete schedule) of all the operations that will be processed over the whole horizon does not make sense. On the other hand, 'sequencing' may be regarded as a relevant decision at the planning level, whereas the exact start times may be considered as a 'detail' that can be fixed 'later' as time passes and disturbances occur.

In other words, and at least in the context of production by lots, the sequencing decisions cannot be considered as 'less important' than the lot-sizing decisions, or a 'detail' to fix after the lot-sizes are determined. In this paper, some alternative approaches are presented, and we show that an integrated approach that we have proposed and extended, tries to overcome the inconsistency between planning and scheduling decisions. For instance, basic principles of this approach have been implemented in the software *Finity* of the Australia-based company *QED International* (see Dye, 1998).

2. The problem

We consider the production of a set of N different items (finished and/or semi-finished) in a general multi-stage system, where the subset of finished products or end items is denoted $\mathcal{N}_0$. Production is carried out in one or several shops, that can be located on one or several production sites. Planning is performed on a horizon of T periods, and the objective is to determine a production plan, i.e. production quantities at every period, that optimizes a given economic criterion, usually the minimization of the sum of production, inventory, and backlog costs. Moreover, we would like the production plan to be *feasible*, i.e., production quantities, that correspond to lots (or jobs in scheduling theory) sent in the shop floor, need to be completed by the end of their associated period. Let (i, l) denote the lot (job) of item i that needs to be completed before the end of period l. The following notation will be used.

Variables:

X_{il}: quantity of item i available at the end of period l.

I^+_{il}: positive inventory level (surplus) of item i at the end of period l.

I^-_{il}: negative inventory level (backlog) of item i at the end of period l.

t_o: start time of operation o of job $(i(o), l(o))$.

Parameters:

D_{il}: demand of item i at the end of period l.

$\mathcal{DS}(i)$: set of the direct successors of item i in the gozinto tree.
g_{ij}: gozinto factor, i.e., the number of units of item i required to produce one unit of item j ($g_{ij} = 0$ if $j \notin \mathcal{DS}(i)$).
c_i^p: production cost per unit of item i.
c_i^{inv}: inventory cost per unit of item i in a period.
c_i^{back}: backlog cost per unit of item i at the end of a period.
c_l: length of period l (avalilable capacity).
$\mathcal{O}$: set of operations.
$\mathcal{A}$: set of pairs of operations in the routings of the products.
($(o, o') \in \mathcal{A}$ means that operation o precedes operation o' in the routing).
$\mathcal{L}$: set of last operations in the routings.
$\mathcal{F}$: set of first operations in the routings.
$i(o)$: item associated with operation o.
$m(o)$: resource on which operation o has to be performed.
$l(o)$: period associated with operation o.
p_o^u: processing time of operation o per unit of item $i(o)$.
L_i: lead time of item i.
$\mathcal{E}$: set of pairs of operations that need to be performed on the same machine.
($(o, o') \in \mathcal{E}$ means that $m(o) = m(o')$).
$\mathcal{S}(y)$: sequence of operations associated with the sequence y
($(o, o') \in \mathcal{S}(y)$ means that o precedes o' in the sequence of a resource).

For the sake of simplicity, we suppose that only end items have external demands.

3. Standard planning and scheduling approaches

In the standard hierarchical approach, the planning decision level first determines an aggregate production plan and then a master production schedule (MPS), i.e. quantities of finished products to produce for every item and every period of some (mid-term) time-discretized horizon. It may happen that the MPS is built on a different (shorter) discretized horizon.

The usual MPS models are described in many textbooks. For instance, a typical capacitated lot-sizing model is the following capacited lot-sizing problem (CLSP) (cf. Salomon, 1991).

$$\begin{cases} \min \displaystyle\sum_{i=1}^{N}\sum_{t=1}^{T} s_i Y_{it} + c_i^{inv} X_{it}^{+} + c_i^p X_{it} & & (1) \\ I_{it-1}^{+} + X_{it} - D_{it} = I_{it}^{+} & \forall i,\ t & (2) \\ \displaystyle\sum_{i=1}^{N} b_i X_{it} \leqslant c_t & \forall t & (3) \\ X_{it} \leqslant \left(\displaystyle\sum_{k=1}^{T} D_{ik}\right) Y_{it} & \forall i,\ t & (4) \\ X_{it},\ I_{it}^{+} \geqslant 0 & \forall i,\ t & (5) \\ Y_{it} \in \{0,\ 1\} & \forall i,\ t & (6) \end{cases}$$

where Y_{it} is a Boolean variable equal to 1 if production of item i takes place in period t (i.e. if $X_{it} > 0$), s_i is the setup cost of item i, and b_i is the per unit of item i capacity absorption. Various extensions (with backlogging, setup times, and so on) can be found in e.g. Salomon (1991).

Once computed, this MPS becomes an input of some MRP-like procedure that translates the MPS into planned orders with a release date and a due date, which in turn becomes an input to some scheduling module (perhaps on a shorter time horizon).

As already mentioned, the sequencing of elementary operations on the machines is ignored, i.e. it is not considered as a decision of the same level of importance as the lot-sizing decision. A 'rationale' for doing so is typically that *if the duration of a period is, say, one week, one should ignore the 'detail' of the time spent by one item on some machine if this time is 'negligible', say, a few minutes. For capacity restriction, only the total workload on the machine needs to be considered at this stage.* Also, a convincing argument is that, in view of the many disturbances that occur in the production system, it is useless and even unrealistic to determine in advance an exact schedule for all the operations, e.g. to determine on Monday morning of the first week of the horizon, the exact start time of an elementary operation that 'should' take place on Wednesday of the third week at 9:30 am.

Another explanation behind this division between planning or lot-sizing problems and scheduling problems is that the research community tends to be divided between researchers investigating planning or lot-sizing problems, mostly based on the use of linear programming involving continuous variables and some binary variables modeling set-ups, and researchers interested in scheduling problems, that are pure combinatorial optimization problems.

In recent years, some effort has been made to bridge the gap between the two research communities. A new class of scheduling problems have emerged by considering that a job is a lot, and can be divided into sublots (see for instance Potts and Van Wassenhove, 1992; Crauwels, Potts, and Van Wassenhove, 1997; Hariri and Potts, 1997; Brucker, Gladky, Hoogeveen, Kovalyov, Potts, Tautenhahn, and Van de Velde, 1998) like the lot streaming problem (see for instance Baker and Jia, 1993; Dauzère-Pérès and Lasserre, 1997; Glass, Gupta, and Potts, 1994; Trietsch and Baker, 1993; or Vickson, 1995). Researchers working in lot-sizing have started to incorporate more realistic scheduling constraints in their model (see Salomon 1991; Kuik, Salomon, and Van Wassenhove, 1994; or Fleischmann and Meyr, 1997). Drexl and Kimms (1997) survey some recent advances on different problems like the DLSP (discrete lot-sizing and scheduling problem), CSLP (continuous setup lot-sizing problem), PLSP (proportional lot-sizing and scheduling problem), and GLSP (general lot-sizing and scheduling problem). As noted by the authors, these problems consider only a single machine, i.e., one production stage. They also use small time buckets, which usually very quickly leads to intractable problems when the number of items or the period length increases. An important drawback of these approaches, if one wants to implement them in a real-world setting, is that they are monolithic. They assume that both planning and scheduling decisions are taken at the same level, and will be implemented as such. However, as already discussed, lot sizes will often be sent as input to the scheduling level, which has its own internal decision procedures (often manual). These procedures might not be optimal, and will have to integrate specific constraints ignored in the models discussed in Drexl and Kimms (1997) (routing flexibility, multi-resource, and so on). Hence, 'optimal' production plans will actually be unfeasible.

Therefore, in most capacitated lot-sizing models, the 'aggregate' capacity constraint (3) is considered to be enough (necessary and sufficient). It states that the total workload on the machines is less than the capacity available on the machines. For this to be true, an implicit assumption has to hold. This assumption is that the *transfer lot* is *very small* (in fact, ideally, 'infinitesimal'). When it is true, the time spent by one item on a machine is almost negligible and it goes immediately on the next machine in the routing. The machines are treated as 'parallel'. Thus, an abstract model with infinitesimal transfer lots, is in fact a 'fluid' model with the inventory balance difference equations replaced by an ordinary differential equation $dI_{it}/dt = x(t) - d(t)$, with x and d now being 'rates' of production and demand respectively. The 'instantaneous' constraint capacity on machine m is just $\sum_i b_{im} x_{it} \leqslant 1$. For more details on such production models, the reader is referred to Gershwin (1994) (see also Sharifnia, 1994; Weiss, 1996).

However, this *small transfer lot assumption* is *not* satisfied in many practical situations where the production is by *lots*, i.e. when the lot is an indivisible entity. Indeed, when the production is by lots of significant size, the time spent on each machine by a lot (not an item) is not negligible compared to the total time the machines are available in one production period (a day for instance). If the routing of a product consists of n elementary operations, with p_i^u being the per-unit processing time on a machine at stage i, the total time spent on one machine by a lot of q items is $q \times p_i^u$, and the total minimum time spent by a lot in the workshop is just $q\sum_i p_i^u$.

Example: Consider the production of two items A and B in a flowshop (with no setup time) over a discretized horizon. The routing consists of three machines M_1, M_2, and M_3. The per-unit processing times are as follows: $p_{A1}^u = 2$, $p_{A2}^u = 1$, $p_{A3}^U = 1$, $p_{B1}^u = 3$, $p_{B2}^u = 1$, and $p_{B3}^u = 2$. Then, the production quantities (lots) at some period t of 60 time units, X_{At} and X_{Bt}, should satisfy the capacity constraints (3):

$$2X_A + 3X_B \leqslant 60,\ X_A + X_B \leqslant 60,\ X_A + 2X_B \leqslant 60$$

on each of the three machines M_1, M_2, and M_3, respectively.

For instance, let X_{At} andX_{Bt} be equal to 10. The capacity constraints are not even saturated, so that one may think that some safety capacity is left. The per-unit capacity consumption coefficients, i.e. 1, 2, and 3 time units are rather small compared to the 60 time units available in period t. In isolation, a unit of item A can be completed in four time units, whereas one unit of item B requires 6 time units. Suppose now that, in period t, one sequences the lot of item A before the lot of item B, and no preemption is allowed. It takes at least 80 time units to complete both lots (see Fig. 1). This schedule

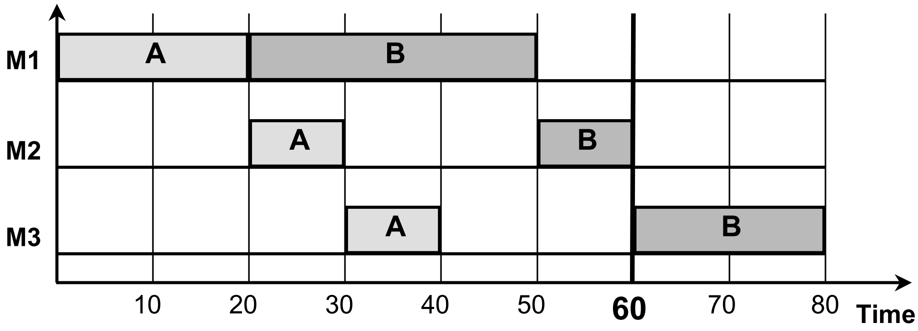

Fig. 1. Item A sequenced before item B.

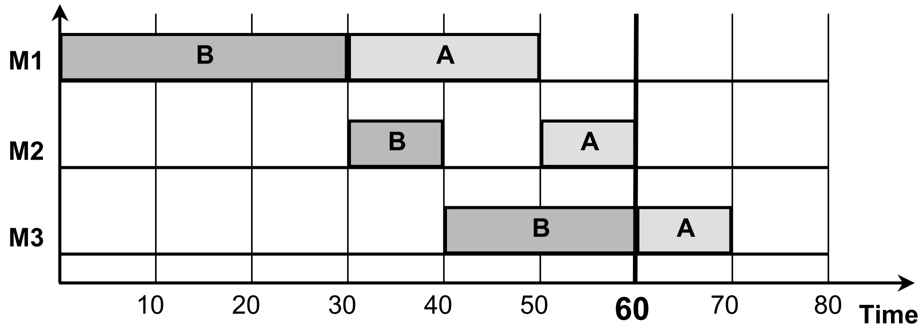

Fig. 2. Item *B* sequenced before item *A*.

can be improved, and the completion time decreased to 70 time units, by sequencing the lot of item *B* before the lot of item *A* (see Fig. 2). However, in both cases, the completion time is well above the 60 time units that are available in the period.

Our claim is that, in this case, the sequencing of operations on the machines cannot be ignored when defining the lot sizes. Indeed, the exact capacity restriction (i.e. the constraints on the lot sizes) depends very much on the sequencing. There is a complex interplay between lot-sizing and sequencing to achieve a good makespan. Different studies have demonstrated the impact of lot sizes on the makespan (e.g. cf. Karmarkar, 1987).

This is why the MRP-like procedures (originally uncapacitated) are not satisfactory. The more sophisticated procedures like MRP-II, even if they partly recognize this fact, are still far from being satisfactory. Indeed, a more than questionable concept in those MRP procedures is the notion of *lead time* considered as an external input data, whereas it is precisely a consequence of the scheduling procedure that will be used (cf. several remarks in Baker, 1993). The reader is also referred to Karmarkar (1987) for an analysis of the impact of lot-sizing on lead times and Karmarkar (1993) (and the references therein) for the discussion on new research directions on models incorporating lead times.

Being procedures that do not (and do not want to) consider the sequencing of operations since it occurs *before* any scheduling decision has been taken, those MRP procedures suffer from a fundamental misunderstanding. The lead time concept is a 'detour' that has been introduced partly to avoid considering the sequencing of operations.

4. Some alternatives

4.1. Underestimating the aggregate capacity

In this approach, one still uses a classical lot-sizing model as in Section 2, but one deliberately underestimates c_t in (3), i.e. the amount of time where the machines are available. The overlooked impact of sequencing is compensated by replacing, in (3), c_t by αc_t for some scalar $\alpha (0 < \alpha < 1)$

modeling the capacity lost through sequencing. However, α may be much less than 1 before the resulting MPS is feasible. Another serious drawback of this approach is that one supposes in advance that the machines are working at no more than $100 \times (1 - \alpha)\%$ of their capacity. This is a problem because one would like to keep the bottleneck machines as busy as possible.

Practitioners using CAPM (computer-aided production management) softwares, and after a CRP (capacity requirements planning) analysis is performed, advance or postpone production orders proposed by MRP. This is generally done based on machine availability reduced by a given percentage. This percentage is usually attributed to uncertainty in the production system that needs to be accounted for, and not to the sequencing constraints that are predictable.

We came across a similar type of approach in one of the assembly factories of an important car manufacturer. This factory assembles small and medium-sized cars. At the top management level, the overall amount of cars to be produced over the year is decided for each car type. These quantities are first refined on a time horizon of several months, and then allocated to the various factories on a time horizon of several weeks. These quantities correspond to detailed products, i.e. a car type with all its options (color, type of engine, electronic equipment, and so on). In the factory, from experience, they knew that the requested amount of cars to be assembled could not be performed on the line. Hence, they were removing 25% from the requested amount of every finished item. The common belief was that this factor of 25% was a result of randomness in the system, mostly scrap problems induced by re-work. At the final assembly level, optimization software is used to solve the car sequencing problem, i.e. how to sequence cars on the line so that capacity constraints are satisfied. These sequencing constraints are mostly human constraints related to the fact that e.g. given the speed of the assembly line and safety constraints, an operator cannot perform a given type of operation on more than two cars out of five. We could show them that in fact the main reason for their factor of 25% was the combined effect of the sequencing constraints. Although we did not directly participate, they conducted a more thorough study to analyze the impact of the sequencing constraints, and how they could specify a factor on every car type, instead of the common factor they were using. It was not possible to convince them that production quantities and the car sequence had to be determined simultaneously, or at least in a more consistent way.

4.2. An integrated model

In this approach, one directly considers the sequencing decisions while computing the lot sizes. By building an integrated model, one may derive exact (detailed) capacity constraints on the lot sizes. To every 'lot' of size X_{il} corresponds a 'job' J_{il} that has to be completed by the end of period l (but may be started in earlier periods). The lead time L_i can be either strictly positive, which indicates that job J_{il} cannot start before period $l - L_i$, or equal to 0, meaning that no direct constraint is imposed on the start of J_{il} and thus its overall processing time. We shall discuss at the end of this section the difference between how we use lead time, and how it is used in MRP.

Ideally, an integrated model has to consider the sequencing constraints associated with the scheduling problem, and is of the form:

$$\begin{cases}
\min \sum_{i,l} c_i^p X_{il} + \sum_{i,l} (c_i^{inv} I_{il}^+ + c_i^{back} I_{il}^-) & & (7) \\
I_{il}^+ - I_{il-1}^- = I_{il-1}^+ - I_{il-1}^- + X_{il} - \sum_{j \in \mathscr{DS}(i)} g_{ij} X_{jl+L_j} - D_{il} & \forall i, \forall l & (8) \\
t_{o'} \geqslant t_o + p_o^u X_{i(o)l(o)} & \forall (o, o') \in \mathscr{A} & (9) \\
t_{o'} \geqslant t_o + p_o^u X_{i(o)l(o)} & & \\
\text{or} & & \\
t_o \geqslant t_{o'} + p_{o'}^u X_{i(o')l(o')} & \forall (o, o') \in \mathscr{S}(y) & (10) \\
t_o + p_o^u X_{i(o)l(o)} \leqslant \sum_{l=1}^{l(o)} c_l & \forall o \in \mathscr{L} & (11) \\
t_o + p_o^u X_{i(o)l(o)} \geqslant \sum_{l=1}^{l(o)-1} c_l & \forall o \in \mathscr{L} & (12) \\
t_o \geqslant \sum_{l=1}^{l(o)-L_{i(o)}} c_l & \forall o \in \mathscr{F} \text{ such that } L_{i(o)} > 0 & (13) \\
X_{il}, X_{il}^-, I_{il}^+, I_{il}^- \geqslant 0 & \forall i, l & (14) \\
t_o \geqslant 0 & \forall o & (15)
\end{cases}$$

Constraint (8) is the classical inventory balance equation. Constraint (9) is the conjunctive constraint between operations in the routings, and constraint (10) is the disjunctive constraint between operations that have to be sequenced on the resources. Constraint (11) makes sure that the production quantity X_{il} is completed before the end of period l, and constraint (12) that it is not completed before the start of period l. Finally, if a lead time is imposed on item i, constraint (13) guarantees that production of X_{il} is performed between periods $l - L_i$ and l, and does not start before.

When the lead time L_i of a given item i is strictly positive, then the availability of the necessary components, when starting production of X_{il}, is ensured through the inventory balance equation (8) and the capacity constraint (11). On the other hand, when $L_i = 0$, it becomes necessary to complete lots X_{jl}, of items j such that $g_{ji} > 0$, before lot X_{il} in order to ensure that the necessary components are available. This is done through the routing constraints (9). Namely, the following constraint needs to be considered:

$$t_{o'} \geqslant t_o + p_o^u X_{i(o)l(o)} \quad \forall (o, o') \text{ such that } o \in \mathscr{L},\ o' \in \mathscr{F},\ g_{i(o)i(o')} > 0 \text{ and } L_{i(o')} = 0$$

which ensures that the first operation of lot $X_{i(o')l}$ starts after the last operations of lots $X_{i(o)l}$ such that $g_{i(o)i(o')} > 0$ (i.e. $i(o)$ is a component of $i(o')$) are completed. This is done by adding the corresponding pairs of operations (o, o') in the set $\mathscr{A}$. This information needs to be fed into the scheduling module of the iterative procedure described in the sequel.

Note that constraint (11) is not necessary for item i if it is the component of an item j (i.e. $g_{ij} > 0$)

such that $L_j = 0$, since the constraint is redundant with constraint (9) between i and j, and constraint (11) on j.

Note that, in our model, the lead time is used in a very different way than in MRP. It does not correspond to an estimated time of production for an item, but it acts as a capacity constraint since constraints (11) and (13) make sure that processing of the lot of an item will not last more than its lead time. The drawback is that the corresponding number of periods is 'reserved', even if production lasts less than the lead time. In our model, lead times can be useful when one wants to follow closely the inventory of semi-finished products, since the time windows in which they are produced are perfectly defined.

5. The iterative procedure

Because of the disjunctive constraints (10), it is very difficult to solve the above problem, even for small-size instances. However, note that, for a *fixed* sequence of operations on the machine, constraint (10) simplifies and reduces to simple linear constraints, as (9). This observation led to an iterative procedure that alternates between two independent modules (see Fig. 3).

- The lot-sizing module that solves the model for a *fixed sequence* y of operations on the machines. Hence, the optimal production plan is computed for the sequence y. When setup times or costs can be ignored, or are always counted, it is a 'simple' linear programming problem, for which many efficient standard packages (OSL, CPLEX, XPRESS-MP, and so on) are available and can solve very large instances.
- The scheduling module that solves the scheduling problem for *fixed sizes* of the lots. With the production plan $X(y)$ computed in the planning module, there is at least a capacity constraint that is tight (otherwise, the plan is globally optimal and the procedure stops), i.e. the last operation of a lot of an item of a period l ends exactly at the end of period l. To improve the production plan, we need to find a sequence y' better than y, i.e. such that all jobs end on time and that operation ends strictly

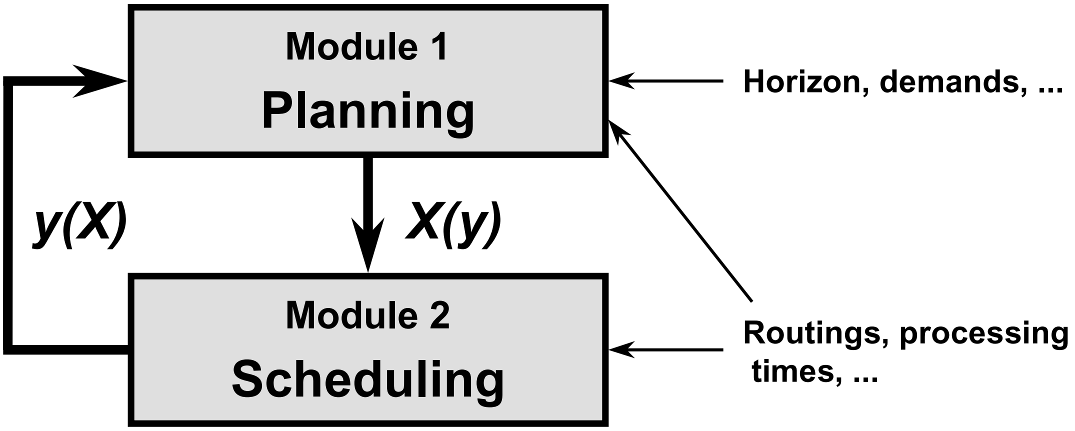

Fig. 3. The iterative procedure.

before the end of the period l. Some time will be available to produce more in l. This problem is equivalent to a scheduling problem with the makespan criterion (see Dauzère-Pérès and Lasserre, 1994).

This procedure has been tested on a sample of test problems (Dauzère-Pérès and Lasserre, 1994). It has been extended to production environments more complex than the job-shop (assembly, multi-stage, multi-site, etc.) . . . (Roux, 1997; Roux, Dauzère-Pérès and Lasserre, 1999).

The procedure has some attractive features:

- it is numerically robust when starting from different initial solutions;
- it provides very good results in a few iterations;
- each module ignores the internal procedure of the other module. The output of the lot-sizing module is the input of the sequencing module and, in turn, the output of the sequencing module (the sequence) is the input of the lot-sizing module (i.e. is used to build the constraints that describe the sequencing of operations on the machines);
- the scheduling procedure can be any scheduling package (a simulation with some priority rules, or any ad-hoc heuristic) and, therefore, can be adapted to the tools used in each particular manufacturing environment.

The efficiency of the overall procedure relies on the efficiency of the sequencing module, for the lot-sizing module is just a linear program. However, as shown in Fig. 4, the procedure is intended to be used at the planning level. The ultimate goal is not to optimally solve the integrated model introduced

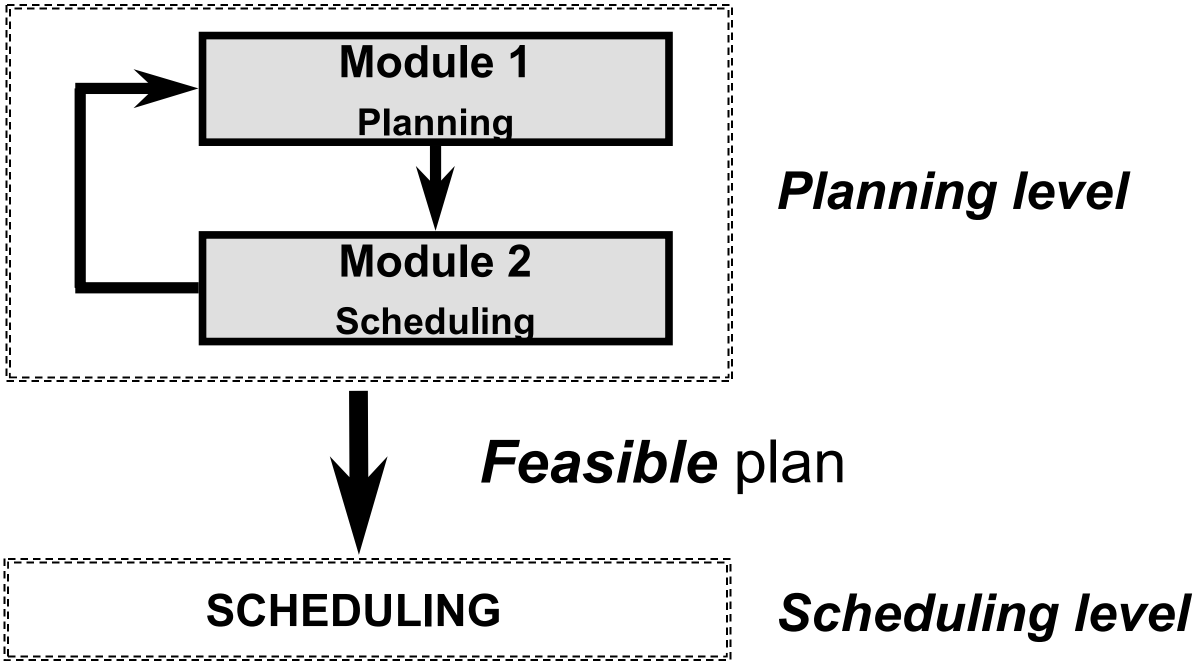

Fig. 4. Using the iterative procedure.

in Section 4.2, i.e. to find an optimal plan *and* schedule, but to determine an optimal *feasible* production plan, considering the capacity of the scheduling level. Hence, to ensure consistency between the planning and scheduling decision levels, it is crucial to ensure the consistency of the scheduling procedures used in the scheduling module of our iterative procedure (and at the planning level), and the ones used at the scheduling level. The scheduling module should represent as closely as possible the actual capacity of the workshop, and the scheduling techniques are part of that capacity. The better the jobs can be sequenced on the shop floor, the higher is the capacity.

Hence, if a sophisticated algorithm is used in the scheduling module, and very simple priority rules are used at the shop floor, the actual capacity will be overestimated in our procedure. Larger quantities than what can actually be produced might be sent to the workshop, and due dates of the jobs will not be met. On the other hand, if simple priority rules are used in the scheduling module, and a very efficient operator schedules the jobs at the scheduling level, the actual capacity will be underestimated in our procedure. The workshop might be idle because not enough quantities are sent, and unnecessary holding and backlog costs will be paid.

However, note that, although similar techniques (or with similar efficiency) should be used in our scheduling module and at the scheduling level, the objectives are different. The scheduling module wants to ensure that the production quantities can be done on time, and the scheduling level might want to consider other qualitative constraints. An efficient way of applying our approach in practice would be to send the schedule determined by the scheduling module to the scheduling level, which could modify it according to its internal objectives and constraints. In addition, as mentioned in the introduction, if the sequencing is important for capacity considerations, the exact schedule determined in the scheduling module is just an indication valid at the time the algorithm is run. Even minor disturbances will make this schedule unrealistic. However, if the disturbances are indeed minor, implementing the ‘sequence’ may be realistic, the exact start times being adjusted as time passes.

Our approach naturally encompasses the actual capacity of the shop floor, including the efficiency with which scheduling is performed. It does not, as in the solving procedures used for the monolithic models surveyed in Drexl and Kimms (1997), suppose that planning and scheduling decisions are taken simultaneously, or that lots are optimally scheduled in the shop floor.

Another important remark which supports the validity of our approach is that only one iteration of our procedure is necessary to outperform the standard hierarchical approach. This is because, by plugging the sequence determined at the scheduling module (and that *would* be used in the shop floor) into our integrated model and running the latter, an at least equally good or better production plan is obtained.

In our implementation, the procedure used in the scheduling module is based on a rather sophisticated algorithm proposed in Dauzère-Pérès and Paulli (1997) and extended in Dauzère-Pérès, Roux and Lasserre (1998). The interested reader is referred to Roux (1997) for more details.

Our solving procedure has been implemented in a decision support system developed on Borland C++ Builder, where linear programming problems are solved using the IBM OSL version 2 library. This DSS, developed at Ecole des Mines de Nantes, allows the various data to be entered in a user-friendly way: products, resources, bill-of-materials, routings, costs, demands, and so on. Various multi-period scheduling policies can be selected (period by period, semi-global or global, see Dauzère-Pérès and Lasserre, 1994), together with the maximum number of iterations, before starting the iterative procedure. The resulting production plan and associated inventory levels can be vizualized on a figure. The associated schedule can also be seen and modified on a Gantt chart.

6. Conclusion

We have tried in this paper to show the limitations of traditional approaches to production planning and scheduling. Because only aggregate capacity constraints are taken into account, decisions taken at the planning level are often inconsistent with the scheduling decisions. A more recent trend (see Drexl and Kimms, 1997) consists in incorporating more detailed and thus more exact capacity constraints in mathematical programming models used to determine lot sizes. However, because solving these models means determining both an optimal production plan and an optimal production schedule, they fail to capture the scheduling performance of the shop floor. Moreover, the rapidly increasing complexity of such models limits the size of the problems that can be solved. In practice, planning and scheduling decisions are still often taken independently. Therefore, we think that planning models should incorporate considerations on how scheduling is performed on the shop floor, rather than assume that planning and scheduling can be done simultaneously at the same decision level.

The approach we propose is an attempt to overcome drawbacks of previous approaches. Our two-step iterative procedure can handle very complex multi-stage manufacturing environments, by leaving the complexity of scheduling to a specific module. Moreover, this module does not have to (and often will not) determine the optimal schedule, but should reflect how sequencing is performed at the scheduling level. Consistency of the production plan is then ensured. Research on our approach is clearly far from being completed. In particular, specialized procedures to handle set-up costs need to be developed.

Acknowledgments

We thank William Roux for helping us clarifying our ideas through many discussions, and Atle Nordli for his helpful comments. This research was partly supported by the CRITT Pays de la Loire Productique. It was performed while the first author was on leave at the Department of Finance and Management Science, Norwegian School of Economics and Business Administration, Helleveien 30, N-5035 Bergen-Sandviken, Norway.

References

Baker, K.R., 1993. Requirement Planning. In: S.C. Graves, A.H.G. Rinnooy Kan, P.H. Zipkin (Eds.), *Logistics of Production and Inventory*, Handbook in Operational Research and Management Science, vol 4, North-Holland, Amsterdam.

Baker, K.R., Jia, D., 1993. A comparative study of lot streaming procedures, *OMEGA* 21, 5, 561–566.

Bitran, G., Tirupati, D., 1993. Hierarchical Production Planning. In S.C. Graves, A.H.G. Rinnooy Kan, P.H. Zipkin (Eds.), *Logistics of Production and Inventory*, Handbook in Operational Research and Management Science, vol 4, North-Holland, Amsterdam.

Brucker, P., Gladky, A., Hoogeveen, H., Kovalyov, M.Y., Potts, C.N., Tautenhahn, T., Van de Velde, S.L., 1998. Scheduling a batching machine, *Journal of Scheduling* 1, 1, 31–54.

Crauwels, H.A.J., Potts, C.N., Van Wassenhove, L.N., 1997. Local search heuristics for single machine scheduling with batch set-up times to minimize total weighted completion time, *Annals of Operations Research* 70, 261–279.

Dauzère-Pérès, S., Lasserre, J.B., 1994. *An Integrated Approach in Production Planning and Scheduling*, Lecture Notes in Economics and Mathematical Systems. Springer-Verlag, Heidelberg.

Dauzère-Pérès, S., Lasserre, J.B., 1997. Lotstreaming in job-shop scheduling, *Operations Research* 45, 4, 584–595.

Dauzère-Pérès, S., Paulli, J., 1997. An integrated approach for modeling and solving the general multiprocessor job-shop scheduling problem using tabu search, *Annals of Operations Research* 70, 281–306.

Dauzère-Pérès, S., Roux, W., Lasserre, J.B., 1998. Multi-Resource Shop Scheduling with Resource Flexibility, *European Journal of Operational Research* 107, 2, 289–305.

Drexl, A., Kimms, A., 1997. Lot sizing and scheduling – Survey and extensions, *European Journal of Operational Research* 99, 221–235.

Dye, R., 1998. Finity: An integrated solution to production planning and scheduling in process manufacturing industries, Technical report, QED International, Australia. http://www.ilog.fr/products/optimization/tech/custpapers/qed.pdf

Fleischmann, B., Meyr, H., 1997. The general lotsizing and scheduling problem, *OR Spektrum* 19, 1, 11–21.

Gershwin, S.B., 1994. *Manufacturing Systems Engineering*. Prentice Hall, Englewood Cliffs.

Glass, C.A., Gupta, J.N.D., Potts, C.N., 1994. Lot streaming in three-stage production processes, *European Journal of Operational Research* 75, 2, 378–394.

Hariri, A.M.A., Potts, C.N., 1997. Single machine scheduling with batch set-up to minimize maximum lateness, *Annals of Operations Research* 70, 1, 75–92.

Karmarkar, U.S., 1987. Lot sizes, lead times and in-process inventories, *Management Science* 33, 409–418.

Karmarkar, U.S., 1993. Manufacturing Lead Times, in S.C. Graves, A.H.G. Rinnooy Kan, P.H. Zipkin (eds): *Logistics of Production and Inventory*, Handbook in Operational Research and Management Science, vol 4, North-Holland, Amsterdam.

Kuik, R., Salomon, M., Van Wassenhove, L.N., 1994. Batching decisions – Structure and models, *European Journal of Operational Research* 75, 2, 243–263.

Lasserre, J.B., 1992. An integrated model for job-shop planning and scheduling, *Management Science* 38, 1201–1211.

Lenstra, J.K., Rinnoy Kan, A.G.H., 1977. New directions in scheduling theory, *Operations Research Letters* 6, 255–259.

Potts, C.N., Van Wassenhove, L.N., 1992. Integrating scheduling with batching and lot sizing: A review of algorithms and complexity, *Journal of the Operational Research Society* 43, 5, 395–406.

Roux, W., 1997. *Une approche cohérente pour la planification et l'ordonnancement de systèmes de prodution complexes*, Thèse en Informatique Industrielle de l'Université Paul Sabatier (Toulouse III), LAAS Report No 97248.

Roux, W., Dauzère-Pérès, S., Lasserre, J.B., 1999. Planning and Scheduling in a Multi-Site Environment. *Production Planning and Control* 10, 1, 19–28.

Salomon, M., 1991. *Deterministic Lotsizing Models for Production Planning*, Lecture Notes in Economics and Mathematical Systems. Springer-Verlag, Heidelberg.

Seyed, J., 1996. Optimal production planning, *OR/MS Today*, 23(2), 56–59.

Seyed, J., 1998. Strenthening key links, *OR/MS Today*, 25(2), 58–61.

Sharifnia, A., 1994. Stability and performance of distributed production control methods based on continuous flow models. *IEEE Transactions on Automatic Control* 39, 4, 725–737.

Smith, D.J., 1978. Material requirement planning, in A.C. Hax (ed.): *Studies in Operations Management*. North-Holland, Amsterdam.

Thomas, L.J., McClain, J.O., 1993. An Overview of Production Planning, in S.C. Graves, A.H.G. Rinnooy Kan, P.H. Zipkin (eds): *Logistics of Production and Inventory*, Handbook in Operational Research and Management Science, vol 4, North-Holland, Amsterdam.

Trietsch, D., Baker, K.R., 1993. Basic techniques for lot streaming, *Operations Research* 41, 6, 1065–1076.

Uzsoy, R., Chung-Yee Lee, Martin-Vega, L.A., 1994. A review of production planning and scheduling models in the semiconductor industry part II: Shop-floor control, *IIE Transactions* 26, 44–55.

Vickson, R.G., 1995. Optimal lot streaming for multiple products in a two-machine flow shop, *European Journal of Operational Research* 85, 3, 556–575.

Vollmann, T.E., Whybark, D.C., Berry, W.L., 1997. *Manufacturing Planning and Control Systems*, 4th ed., Irwin Professional.

Weiss, G., 1996. Optimal Draining of Fluid Re-Entrant Lines: Some Solved Examples. In: F.P. Kelly, S. Zachary, I. Ziedins (eds), *Stochastic Networks: Theory and Applications*, 19–34.