Chapter 5
Planning Models with Stationary Fixed Lead Times

In this chapter, we present optimization models for order release using exogenous planned lead times that remain constant (stationary) over the planning horizon. We describe the material flow dynamics implied by these models, beginning by assuming lead times that are integer multiples of the underlying planning period. We construct a series of linear programming models for this problem and examine their dual, noting several implications that are inconsistent with insights from the queueing models discussed in Chap. 2. We then extend this approach to consider fractional lead times and a more general formulation where a production order may consume capacity in multiple, not necessarily consecutive, periods.

5.1 Preliminaries

The previous chapters have, we hope, set the stage upon which we propose to address the principal topic of this volume: the development of novel, and hopefully more effective, optimization models to support the goods flow problem faced by the planning level, whose purpose is to coordinate the releases of work across multiple production units in the production system or supply chain to meet demand in the best possible manner. Due to the need to match production with demand, the models must take into account the cycle times, the delay between work being released into the production unit and its emergence as completed product that can meet demand.

We shall refer to the smallest unit of work recognized by the goods flow problem as an *order*. Orders may be of external or internal origin; *external* orders represent a specific quantity of a specific product ordered by a specific customer, while *internal* orders are generated by the PPC system for purposes of production management within the production unit, and thus may represent material destined for several customers, a portion of a larger customer order, or simply material intended to

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_5

replenish inventory positions along the supply chain. For the purposes of the goods flow problem, both can be treated in the same manner, so we will use the term “order” for both.

Following the discussion in Chap. 1, a production unit is an organizational unit whose internal operations are not under the control of the planning level, which is tasked with managing the goods flow problem. A production unit consists of several workcenters with limited capacity, through which each order processed in the production unit follows a specified routing. For exposition we assume the routing to be deterministic, ignoring the possibility of random routing due to causes such as alternative resources or rework. While this is certainly not the most general model that could be presented, it is sufficient to convey the essence of the problems we consider. Hackman and Leachman (1989a, b) and Hackman (1990, 2008) provide a much more general treatment encompassing other modes of production such as resource-constrained project scheduling. Per Chap. 2, the cycle time of a unit of work is a random variable that follows some probability distribution, but can only be observed after the fact. We shall use the term *lead time* to denote an estimate of the cycle time used in planning models for the goods flow problem. The focus of this chapter is on planning models that use constant, exogenous lead times to represent the progress of orders through the production unit. For brevity of exposition, we shall refer to these lead times as *fixed lead times*. In this chapter, we consider the simpler case where the planned lead times associated with a production unit and its workcenters remain constant over time, i.e., do not vary across time periods. The more complex case of time-varying planned lead times is treated in the next chapter.

5.2 A Generic Production Unit

Figure 5.1 illustrates a generic production unit that produces a set J of products $j = 1, \ldots, |J|$, for which it has a queue of orders waiting to be processed that have been released by the planning level, and a finished goods inventory location where finished items are stored. The production process uses a set K of different workcenters $k = 1, \ldots, |K|$, with limited capacity, each of which, per Chap. 2, can be modeled as a queueing system. We denote the set of workcenters used by product j by $K(j)$ and the time required to process a unit of product j on workcenter k as a_{jk}. The set of products requiring a workcenter k will be denoted by $J(k)$. The planning horizon is divided into discrete time periods, which we shall assume without loss of generality to be of equal length Δ, such that period t ends at time $t\Delta$. When it causes no ambiguity, we shall assume the time periods to be of unit length so that period t ends at time t. The basic sequence of events taking place in the production unit is as follows:

1. The planning level authorizes the release of an order consisting of a specific quantity R_{jt} of product j to the production unit at time t.
2. The order is released for production and enters the queue for the first workcenter in its routing. Control over its progress through the production unit is transferred to the internal management of the production unit. Upon completion of its

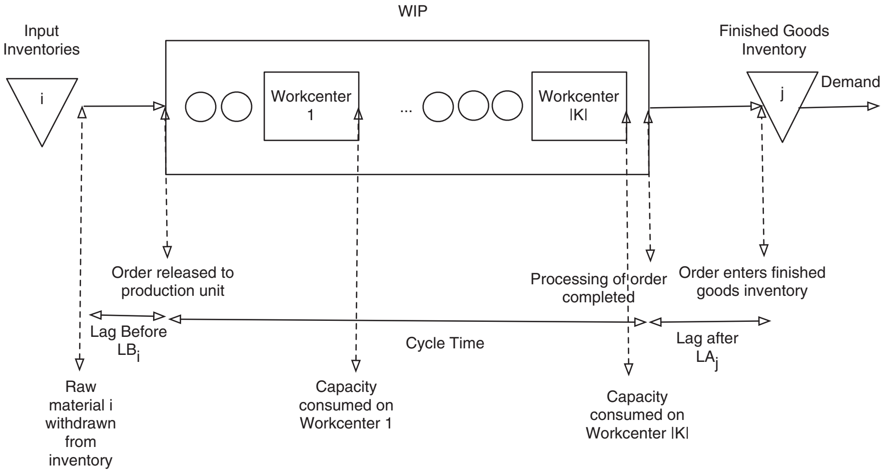

WIP
Input Inventories
i
Workcenter 1
...
Workcenter |K|
Finished Goods Inventory
j
Demand
Order released to production unit
Processing of order completed
Order enters finished goods inventory
Lag Before LB$_i$
Cycle Time
Lag after LA$_j$
Raw material i withdrawn from inventory
Capacity consumed on Workcenter 1
Capacity consumed on Workcenter |K|

Fig. 5.1 Generic Production Unit with Time Lags

processing at each workcenter, the order moves directly to the next workcenter in its routing.

3. The order completes processing on the last workcenter in its routing and moves from the production facility to the production unit's finished inventory location.

Each order of product *j* released in period *t* will wait in the queue for workcenter *k* for an average of Q_{jkt} time units and will require an expected processing time of $P_{jkt} = R_{jt}a_{jk}$ time units, which we assume includes any necessary setup times. The expected cycle time of the order from its release to its entry into finished inventory is thus given by

$$T_{jkt} = \sum_{k \in K(j)} \left(Q_{jkt} + P_{jkt} \right) \tag{5.1}$$

The cycle time of an order at a workcenter *k* is thus the sum of its processing time and its queue time. Per Chap. 2, the queue time is a random variable whose probability distribution depends on the utilization of the workcenter, which is determined by the work release decisions R_{jt}, while the service time is also a random variable. These random variables are represented in (5.1) by their expectations. The expected cycle time of an order is thus given by the sum of its expected processing and queue times at each workcenter *k* in its routing. In practice, additional delays may be incurred, such as transportation time between workstations, preparation of components and raw materials, or transfer of the finished order to finished inventory, which are also likely to be random variables. A wide variety of domain-specific events may need to be considered, such as the need to allow a specified time for lumber to cure before its use in furniture manufacturing or the need to perform a thin-film deposition step within a specified time of a cleaning step in semiconductor manufacturing. The modeling of fixed delays between such events is discussed at length by Hackman and Leachman (1989b) and Hackman (2008). However, the events shown in Fig. 5.1 are sufficient to account for most cases of interest.

5.3 Lead Times in Models of the Goods Flow Problem

The wide range of planning models using fixed exogenous lead times, including both MRP (Orlicky 1975; Baker 1993; Vollmann et al. 2005) and mathematical programming models (Voss and Woodruff 2006; Missbauer and Uzsoy 2011), all assume that as long as all constraints in the model are satisfied, the production unit will be able to produce its output in a manner consistent with the lead time values specified. Thus the lead times serve the planning level as an anticipation function (Schneeweiss 2003) describing the impact of its release decisions on the output of the production units. We view a lead time L_{jk} as a parameter whose value is an estimate of a suitably high percentile of the order cycle time distribution whose mean T_{jkt} is given by (5.1). Hence under normal conditions any order released to the production unit will enter finished goods inventory within $L_j = \sum_k L_{jk}$ time units of its release with high probability. Under this view the lead time is treated as a delay between the release of an order into the production unit and its completion.

Billington et al. (1983) suggest using only the minimum time required to transfer material between operations without considering queue time or processing time; they argue that delays due to limited capacity will be computed by the planning model itself, which should produce materials ahead of time and hold it in finished inventory until needed to meet demand, ignoring the workload-dependent nature of the queue time Q_{jkt}. These transfer times between operations can be modeled as fixed delays following Hackman and Leachman (1989b) if their duration is significant relative to that of the planning period. Another class of planning models treats the fixed lead time not as a delay, but as a time interval within which the production unit must process the order once it is released. We shall first discuss models that treat lead times as delays and treat this latter view in Sect. 5.6.

5.3.1 Planning Models with Fixed Exogenous Lead Times

The vast majority of the mathematical programming models of interest to this volume approach the goods flow problem faced by the planning level following the early formulations of Modigliani and Hohn (1955), Manne (1957), Hanssmann and Hess (1960), and Holt et al. (1955). A finite time horizon is divided into discrete time periods, usually, but not necessarily, of the same length. Decision variables are associated with each period, and the objective is either to minimize total cost or to maximize total contribution (revenue minus variable costs) over the planning horizon. All quantities are treated as deterministic. Following Hackman and Leachman (1989b), such models require three basic sets of constraints:

1. *Inventory or material balance constraints* for all input and finished inventory points, which coordinate material flows through both space and time. These also enforce the satisfaction of demand, which is treated as a material flow out of the production system to an external demand source.

2. *Capacity constraints*, which model how the production activities capture and consume production resources.
3. *Domain-specific constraints* reflecting the special structure and requirements of the particular production environment being modeled. The structure of these constraints will differ widely based on the specific environment under study and hence will not be discussed in detail. We shall focus on the first two constraint sets, which are critical to the model's ability to accurately reflect the realized behavior of the production system for which the plans are developed.

Two points in time are of particular interest: the point at which the production order actually consumes capacity on the resources required to process it and the time it is completed and can be used to meet demand. Knowledge of the former is necessary to ensure that capacity constraints are not violated over time and of the latter to allow accurate prediction of the amount of material available to meet demand over time.

5.3.2 A Single Production Unit

We begin by considering a production unit modeled as in Fig. 5.1. Since the timing and quantity of order releases constitute the link between planning and detailed scheduling within the production unit, release quantities are the primary decision variables of interest. We shall assume all demand must be met without backlogging; this will allow us to focus on representing the behavior of the production unit. Thus negative inventory levels are not permitted at any inventory location. Material flows within the production unit itself are of interest only to ensure that releases are capacity feasible for all workcenters $k \in K$, and hence the production unit can meet demand within the specified lead times.

5.3.2.1 Single Product, Instantaneous Production, Unlimited Inputs

The simplest model of production, encountered in classical inventory models such as the Economic Order Quantity model and the Wagner-Whitin dynamic lot-sizing model (Zipkin 2000; Hopp and Spearman 2008), is instantaneous production where the quantity ordered at a given point in time becomes available immediately upon production being initiated. In mathematical programming models, this implies that cycle time is negligible relative to the length of the planning period, so that the entire quantity R_t of material released into the system during period t is available to meet demand by the end of that period. The assumption of unlimited inputs implies either instantaneous acquisition or sufficient on-hand inventory of all inputs. Thus inputs will never constrain the ability of the production unit to meet demand, and there is no need to model input inventories. Since we have only a single product, the product subscript j is suppressed.

To ensure consistent material flows over time, we model the finished goods inventory level across periods with the material balance equations

$$I_t = I_{t-1} + X_t - D_t, \quad t = 1,\ldots,T \tag{5.2}$$

where I_t denotes the amount of finished goods on hand at the end of period t, X_t the output of the production unit in period t, and D_t the demand during that period. Under instantaneous production, we have $R_t = X_t$; all materials released into the production unit in a period are converted into output by the end of the period. Denoting the amount of finished goods inventory at the start of the first period (the end of period 0) by I_0, (5.2) can be rewritten as

$$I_t = I_0 + \sum_{\tau=1}^{t} X_\tau - \sum_{\tau=1}^{t} D_\tau, \quad t = 1,\ldots,T \tag{5.3}$$

by summing the constraints (5.2) for consecutive periods 1, …, t.

The most common capacity constraint encountered in the literature seeks to ensure that the total production X_t for a given period t, and hence the planned releases R_t, cannot exceed the available capacity C_{kt} of any workcenter k. Since we produce a single product, C_{kt} can be expressed in units of the end item, allowing this constraint to be written as:

$$R_t \le C_{kt}, \quad t = 1,\ldots,T; \quad k = 1,\ldots,K \tag{5.4}$$

Taken together, (5.2) and (5.4) imply that as long as releases do not violate capacity constraints on any workcenter, materials released in period t will be available to meet demand by the end of the same period. If demand D_t in any period t exceeds the capacity of some workcenter k, the only course open to the model is to produce the excess demand in an earlier period $s < t$, holding finished inventory in the periods s to t. Combining (5.3) and (5.4) yields

$$\sum_{\tau=1}^{t} C_{k\tau} \ge \sum_{\tau=1}^{t} X_\tau = \sum_{\tau=1}^{t} R_\tau \ge \sum_{\tau=1}^{t} D_\tau - I_0, \quad t = 1,\ldots,T, \quad k \in K, \tag{5.5}$$

as a necessary condition for a feasible solution to exist. The only reason to release an order in advance of the period in which it is due is lack of capacity at some workcenter k in that period. Denoting the unit cost of holding FGI for one period by h_t and the unit incremental cost of production by c_t, the planning model can be written as:

$$\min \sum_{t=1}^{T} \left(h_t I_t + c_t R_t \right) \tag{5.6}$$

subject to

$$I_t = I_{t-1} + R_t - D_t, \quad t = 1,\ldots,T \tag{5.7}$$

$$R_t \le C_{kt}, \quad t = 1,\ldots,T, \quad k \in K \tag{5.8}$$

$$I_t, R_t \geq 0, \quad t = 1, \ldots, T \tag{5.9}$$

This model, although simplistic in its assumptions, has all the basic components of a production planning model: decision variables associated with each period (the R_t), state variables arising from the decision variables and the constraints (the I_t), an objective function minimizing the sum of production and inventory holding costs (5.6), material balance constraints (5.7) for the finished inventory location, and capacity constraints (5.8) for each resource k in each period t.

The capacity constraint (5.8) ensures that the total planned resource usage during the planning period does not exceed the amount of the resource available during the period. This is necessary, but not sufficient, to ensure that the planned releases can actually be processed within the planning period, since the model does not control the timing of work arrivals at the workcenter within the period. If for some reason such as a machine failure on the shop floor, 75% of the amount released became available only in the second half of the planning period, the workcenter might well not be able to process all of it by the end of the period.

5.3.2.2 Single Product, Non-instantaneous Production

The model (5.6)–(5.9) is not realistic when the magnitude of the workcenter cycle times $Q_{jkt} + P_{jkt}$ is significant relative to that of the planning period. The most common representation of this situation in the literature is a fixed lead time L representing the estimated time required for work released in a given period to become available to meet demand, most commonly expressed as an integer number of planning periods.

Under these assumptions, material released into the production unit during period t becomes available for use L time periods later during period $t + L$, implying that $X_t = R_{t-L}$. The material balance constraints for the finished inventory are now

$$I_t = I_{t-1} + X_t - D_t = I_{t-1} + R_{t-L} - D_t, \quad t = 1, \ldots, T \tag{5.10}$$

This is exactly the model of lead times used in MRP in its backward scheduling phase, where the fixed lead time represents the amount of time elapsing between the time an order for a BOM item is placed and its receipt (Baker 1993; Voss and Woodruff 2003). Since we have only one product (end item), the product index j remains suppressed.

Under instantaneous production, an order consumes capacity at each resource k in the period in which it is released, rendering constraints (5.8) sufficient to ensure capacity feasible releases. However, when lead times exceed one period a question of timing arises—at what point in the lead time L does the job consume capacity on a given resource k? This requires knowledge of the process routing, the sequence in which the different resources are utilized by the order. Without loss of generality, we shall assume that the order visits each resource exactly once in a known, deterministic sequence and that the resources are indexed in the order of their use. Thus resource

$k = 1$ is the first resource used in the routing, and resource $k = |K|$ the last one before the order enters finished inventory. Let L_k denote the estimated delay between the release of the order to the production unit and its becoming available for processing on workcenter k. Thus L_k represents an estimate of the total cycle time of the order at all workcenters in its routing prior to k, implying that

$$E\left[Q_{kt} + P_{kt}\right] = L_k - L_{k-1} \tag{5.11}$$

Clearly we must have

$$\max_{1 \le k \le |K|} \left\{L_k\right\} \le L \tag{5.12}$$

for consistency. Our capacity constraints (5.8) now take the form

$$R_{t-L_k} \le C_{kt}, \quad \text{for all} \quad k \in K; \quad t = 1,\ldots,T \tag{5.13}$$

Since no inventory is held within the production unit other than the WIP waiting for processing or in transit between stages, the output of individual workcenters is represented to capture their incremental costs of production and their limited capacity in each period. By the definition of the lead times L_k, an order processed on workcenter k in period t will have been released in period $t - L_k$. For simplicity of exposition, we shall assume that the total production cost of an order completed in period t, given by

$$c_t = \sum_{k=1}^{|K|} c_{k.t-\left(L-L_k\right)} \tag{5.14}$$

where c_{kt} denotes the unit cost of production on workcenter k in period t, is assessed in period t; this could easily be relaxed at the expense of additional notation. The single-product multiple workcenter model with integer lead times L_k associated with each resource k, and an overall lead time L associated with the entire production unit, is as follows:

$$\min \sum_{t=1}^{T} \left(h_t I_t + c_t R_{t-L}\right) \tag{5.15}$$

subject to

$$I_t = I_{t-1} + R_{t-L} - D_t, \quad t = 1,\ldots,T \tag{5.16}$$

$$R_{t-L_k} \le C_k, \quad t = 1,\ldots,T, \quad k \in K \tag{5.17}$$

$$I_t,\ R_t \ge 0, \quad t = 1,\ldots,T \tag{5.18}$$

Decision variables with non-positive subscripts correspond to decisions made prior to the start of the planning horizon that are known with certainty, and as such are parameters of the model. This is essentially the step-separated formulation of

Leachman and Carmon (1992), without the alternative production routings considered in that paper. The amount of production that can take place on resource k in a given period t is limited by both the capacity C_{kt} and the amount of work available for processing, given by past releases R_{t-L_k} per (5.17). Hence the amount of WIP available to process on workcenter k in period t is simply R_{t-L_k}. The total amount of WIP in the production unit—the amount of material that has been released but not yet completed—is given by

$$W_t = \sum_{\tau=t-L+1}^{t} R_\tau = \sum_{\tau=t+1}^{t+L} X_\tau \tag{5.19}$$

This quantity does not appear in LP models of production planning, such as those discussed in Johnson and Montgomery (1974), Hackman and Leachman (1989a, b), and Voss and Woodruff (2006) that treat fixed lead times as a delay between order release and completion. The reason for this is apparent from (5.19): when a fixed lead time represents a delay the amount of WIP is determined by the lead time L and the releases R_t; any WIP holding cost can be incorporated into the incremental production cost c_t.

The movement of material through a system with four machines in series under this model is traced in Fig. 5.2. The vertical axis shows the lead times for each machine, and each timeline the material processed by each machine in each period, identified by the period of its release to the first machine, machine 1. The material released in each period is indicated by the numeral above it; thus, material released at the start of period 1 is indicated by a “1” above the line indicating the material. Material released in a given period is shown with a bar of a given color until it exits the system; thus the material released in period 1 is shown as a red bar as it proceeds through the machines. Material released at the start of the planning horizon, at the start of period 1, indicated by the red bars, becomes available to machine 2 at the start of period 2, is in WIP at machine 2 at the start of period 3, is available to

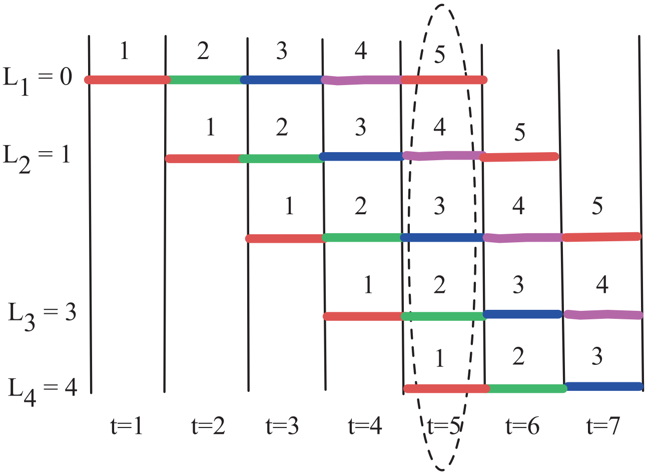


Fig. 5.2 Timing of material flow under fixed lead times

machine 3 at the start of period 4, and is available to machine 4 at the start of period 4. At the end of period 5 or, equivalently, the start of period 6, the WIP at machine 4 consists of the material entering the system (i.e., released to machine 1) in period 2; at machine 3, the material released in periods 3 and 4, and at machine 2, that released in period 5. The figure also indicates that not all the WIP at a machine at the start of a period is necessarily available to be processed at the machine in the period. For example, at the start of period 6, material entering the system in periods 3 and 4 is in WIP at machine 4, but even if the machine has sufficient capacity, only the material entering in period 3 will be processed. In other words, WIP cannot accumulate, but flows through the system in discrete units equal to the quantity released in each period.

The model assumes that WIP will not accumulate in the system over time; only the material released in period $t - L_k$ is available to resource k for processing in period t. Equivalently, all R_t units of product released in period t are assumed to move through the production process as a single entity, occupying capacity on each workcenter within a single period. Since (5.17) ensures that releases do not exceed capacity, the system can always process this quantity in a single period. The remaining WIP still to be processed by the workcenter, given by

$$\tilde{W}_{kt} = \sum_{\tau=t-L_k+1}^{t-(L_k-L_{k-1})+1} R_\tau \tag{5.20}$$

has no effect on the cycle time of the workcenter, which can never exceed $L_k - L_{k-1}$ as long as the capacity C_{kt} of the resource in period t is not exceeded. The lead time L_k simply delays the arrival of work to the workcenter after its release into the production unit; it does not describe the behavior of the workcenter itself.

Examination of constraints (5.16) and (5.17) reveals another consequence of the fixed lead times: the output of the production unit in periods 1 through L cannot be influenced by release decisions in periods 1, …, $L-1$ but is determined by release decisions in periods $-L+1$ through 0 which, since they lie in the past, are assumed to be known with certainty. Thus positive fixed lead times bring the need to initialize the model with information about decisions in the early periods of the planning horizon. These quantities are analogous to the scheduled receipts used in MRP calculations (Baker 1993; Jacobs et al. 2011). Similarly, the model will not plan releases in periods $T - L + 1$ through T, since this material can only meet demand in periods $T + 1$ through $T + L - 1$ that lie outside the planning horizon. Thus the use of fixed lead times requires specifying boundary conditions for the planning models at the beginning and end of the planning horizon.

The timing of releases and output under fixed lead times is illustrated in Fig. 5.3, which assumes a fixed lead time of $L = 2$ periods. Releases R_t in each period t are assumed to be uniformly distributed across the period. Hence the output X_3 in period 3 is determined by the amount of releases R_1 in period 1. However, the output X_1 of the production unit in period 1 lies within the fixed lead time, and hence depends on decisions made in the past, in period $t = -1$. To avoid introducing additional notation for these historical release decisions associated with periods $t = -L + 1$ through $t = 0$, we assume henceforth that any decision variable with a non-positive subscript

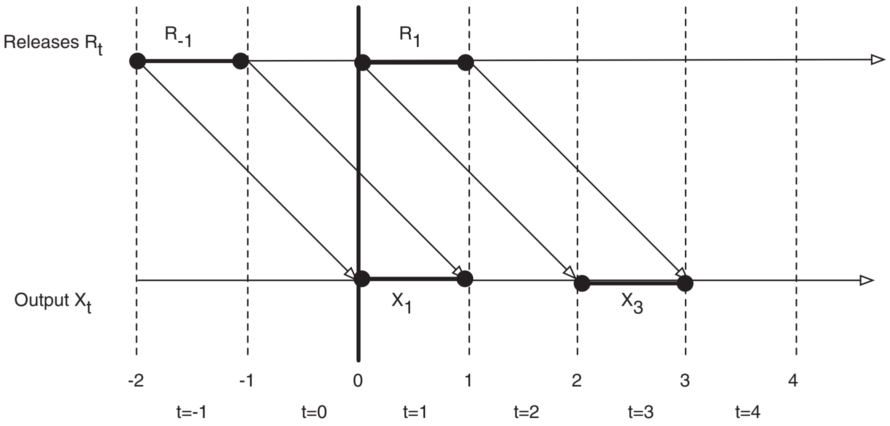


Figure 5.3 Timing of material flows under integer fixed lead times

is a parameter corresponding to a historical decision. Under this model of fixed lead times, the time series X_t, $t = 1,\ldots,T$ representing the output of the workcenter is simply the time series R_t, $t = 1,\ldots, T$ of the releases shifted L periods to the right.

Hence under fixed lead times, the output variables X_t and release variables R_t are completely interchangeable. We have written our formulation in terms of the release variables R_t, but since $X_t = R_{t-L}$ it is straightforward to write it in terms of the output variables X_t.

Finally, the model (5.15)–(5.18) can be rewritten using (5.3) to eliminate the inventory variables. Defining I_0 to be the amount of finished goods inventory on hand at the start of the first period in the planning horizon, we see that

$$\begin{aligned} I_t &= I_0 + \sum_{\tau=1}^{t} R_{\tau-L} - \sum_{\tau=1}^{t} D_\tau \\ &= I_0 + \sum_{\tau=1}^{t} X_\tau - \sum_{\tau=1}^{t} D_\tau \end{aligned} \tag{5.21}$$

Substituting (5.21) into (5.15) yields

$$\begin{aligned} \sum_{t=1}^{T} \left(h_t I_t + c_t X_t\right) &= \sum_{t=1}^{T} \left[h_t \left(I_0 + \sum_{\tau=1}^{t} X_\tau - \sum_{\tau=1}^{t} D_\tau \right) + c_t X_t \right] \\ &= \sum_{t=1}^{T} \left[h_t \sum_{\tau=1}^{t} X_\tau + c_t X_t \right] + \sum_{t=1}^{T} h_t \left[I_0 - \sum_{\tau=1}^{t} D_\tau \right] \\ &= \sum_{t=1}^{T} \left[\sum_{\tau=1}^{t} h_\tau R_{\tau-L} + c_t R_{t-L} \right] + \sum_{t=1}^{T} h_t \left[I_0 - \sum_{\tau=1}^{t} D_\tau \right] \\ &= \sum_{t=1}^{T} \left[\sum_{\tau=t}^{T} h_\tau R_{t-L} + c_t R_{t-L} \right] + \sum_{t=1}^{T} h_t \left[I_0 - \sum_{\tau=1}^{t} D_\tau \right] \end{aligned}$$

Discarding constants independent of the decision variables, we can rewrite (5.15)–(5.18) as

$$\min \sum_{t=1}^{T} \left(\sum_{\tau=t}^{T} h_\tau + c_t \right) R_{t-L} \tag{5.22}$$

$$\sum_{\tau=1}^{t-L+1} R_\tau \geq \sum_{\tau=1}^{t} D_\tau - I_0, t = 1,\ldots,T \tag{5.23}$$

$$R_{t-L_k} \leq C_k, k = 1,\ldots,K; t = 1,\ldots,T \tag{5.24}$$

$$R_t \geq 0, t = 1,\ldots,T \tag{5.25}$$

Model (5.22)–(5.25) shows that the I_t variables are not essential; they are a consequence of the primary decisions, given by the releases R_t, and the constraints describing the behavior of the system. While the model (5.6)–(5.9) is probably more familiar to the reader, as it is widely used in textbooks, the model (5.22)–(5.25) provides some advantages when analyzing the structure of optimal solutions, particularly the dual solutions that we shall examine later in this chapter.

This basic formulation can be extended in a number of directions without materially affecting its structure. Models involving lot-sizing considerations due to the presence of setup costs or setup times, such as that of Billington et al. (1983) or those studied by Pochet and Wolsey (2006), involve integer variables—a significant difference from a computational perspective—but their treatment of capacity and lead times is essentially the same. Far more elaborate objective functions are possible, but our emphasis is on the representation of production capacity and material flow. The assumption of no backlogging can be relaxed in the standard manner (Johnson and Montgomery 1974). Since a backlog corresponds to a negative inventory level, we can represent the net inventory level N_t as the difference of two nonnegative variables $N_t = I_t - B_t$, where I_t denotes on-hand, positive inventory, at the end of period t, and B_t the backlog. Since the column vectors for I_t and B_t in the constraint matrix of the linear programming model will be linearly dependent, both variables cannot take positive values in an optimal solution.

5.3.2.3 Multiple Items

The single-item multiple workcenter model (5.22)–(5.25) extends to the multi-item case with items $j \in J_\mathrm{F}$ in a very natural manner. Capacity constraints at each workcenter k must now consider the total capacity consumption by all items $j \in J(k)$ using that workcenter in each period, and separate finished goods inventory balance equations must be written for each product j. All lead time parameters are now product-dependent, with L_j denoting the lead time of product j from release until completion and L_{jk} its lead time from release until its availability for processing at workcenter k. With these changes, the multi-item model can be written as:

$$\min \sum_{t=1}^{T} \left[\sum_{j \in J_{\mathrm{F}}} \left(h_{jt} I_{jt} + \sum_{k \in K(j)} c_{jkt} R_{j,t-L_k} \right) \right] \tag{5.26}$$

subject to

$$I_{jt} = I_{j,t-1} + R_{t-L_j} - D_{jt}, \quad \forall j \in J_{\mathrm{F}}, \quad t = 1,\ldots,T \tag{5.27}$$

$$\sum_{j \in J(k)} a_{jk} R_{t-L_{jk}} \le C_{kt}, \quad \forall k \in K, \quad t = 1,\ldots,T \tag{5.28}$$

$$I_{jt}, \quad R_{jt} \ge 0, \quad \forall j \in J_{\mathrm{F}}, \quad t = 1,\ldots,T \tag{5.29}$$

The only representation of resource contention between the products j at the workcenters k is the left hand side of (5.28), which is linear in the release quantities of each product. This is in marked contrast to Fig. 2.2, where the output of the resources is a concave non-decreasing function of the workload, determined by the production quantities. The presence of multiple products with different processing times on the workcenter will result in increased coefficients of variation of the processing times P_{jkt} and a downward shift in the output function. The lead times L_{jk} are also unaffected by production quantities, in contrast to the highly nonlinear behavior of the cycle time with workload seen in Fig. 2.1. It begins to be apparent that the workcenter behavior described by this model differs quite fundamentally from that of the queueing models discussed in Chap. 2. The inventory variables can also be eliminated using (5.21), resulting in the formulation

$$\min \sum_{j \in J} \sum_{t=L_j+1}^{T} \left(\sum_{\tau=t}^{T} h_j \right) R_{j,t-L_j} = \sum_{j \in J} \sum_{t=L_j+1}^{T} \left(\left(T - t + 1 \right) h_j \right) R_{j,t-L_j} \tag{5.30}$$

subject to

$$\sum_{\tau=1}^{t} R_{j,\tau-L_j} \ge \sum_{\tau=1}^{t} D_{j\tau} - I_{j0}, \quad \forall j \in J, \quad t = 1,\ldots,T \tag{5.31}$$

$$\sum_{j \in J(k)} a_{jk} R_{t-L_{jk}} \le C_{kt}, \quad \forall k \in K, \quad t = 1,\ldots,T \tag{5.32}$$

$$R_{jt} \ge 0, \quad \forall j \in J, \quad t = L_j,\ldots,T - L_j \tag{5.33}$$

We now use this formulation to discuss the dual model and its interpretation.

5.4 Dual Formulation

Unless it is infeasible or its optimal value is unbounded, any linear program is associated with another linear program, its dual, whose optimal value is equal to that of the original (the primal) at optimality (Bazaraa et al. 2004). Each decision variable in the dual is associated with a constraint in the primal and each dual constraint with a primal decision variable. Thus, the generic linear program

$$\min \sum_{j=1}^{n} c_j x_j \tag{5.34}$$

subject to

$$\sum_{j=1}^{n} a_{ij} x_j \geq b_i, \quad i = 1,\ldots,m \tag{5.35}$$

$$x_j \geq 0, \quad j = 1,\ldots,n \tag{5.36}$$

will be associated with its dual

$$\max \sum_{i=1}^{m} b_i y_i \tag{5.37}$$

subject to

$$\sum_{i=1}^{m} a_{ji} y_i \leq c_j, \quad j = 1,\ldots,n \tag{5.38}$$

$$y_i \geq 0, \quad i = 1,\ldots,m \tag{5.39}$$

The dual variables y_i associated with each primal constraint i correspond to the Lagrange multipliers associated with that constraint, representing the partial derivative of the optimal objective function value with respect to the right-hand side b_i of constraint i at optimality. An important property arising from the Kuhn–Tucker optimality conditions for linear programs is the complementary slackness condition

$$y_i \left(\sum_{j=1}^{n} a_{ij} x_j - b_i \right) = 0, \quad i = 1,\ldots,m \tag{5.40}$$

The dual variables have an economic interpretation that is often helpful in interpreting the results of a model. An important advantage of the models developed in Chap. 7 is their ability to provide richer dual information than that obtained from the models discussed in this chapter.

Since our primary concern lies with production planning models, we discuss duality in an intuitive, heuristic fashion; rigorous mathematical treatments are given by Bazaraa et al. (2004) and Bertsimas and Tsitsiklis (1997). Correct interpretation of dual variables can be quite subtle, especially in the presence of a degenerate optimal solution where some constraints are redundant; extensive discussions of these issues are given by Jansen et al. (1997), Koltai and Terlaky (2000), and Rubin and Wagner (1990). To avoid the extensive mathematical digressions required to address the issues in estimating dual prices in the face of degenerate optimal solutions, our discussion will assume that all optimal solutions are non-degenerate, closely following the development in Kefeli (2011) but omitting some details to focus on insights.

We will develop the dual formulation for the model (5.30)–(5.33). For further simplicity in exposition, we shall assume all costs are time-stationary such that, for example, $c_{jt} = c_j$ for all periods t. In this case, the no-backlogging assumption implies that as long as a feasible solution exists, in any optimal solution total production of any product will exactly equal its total demand net of the initial inventories I_{j0}, and the production costs will have no influence on the optimal solution. This results in the simplified primal linear program

$$\min \sum_{j\in J} \sum_{t=L_j+1}^{T} \left(\sum_{\tau=t}^{T} h_j \right) R_{j,t-L_j} = \sum_{j\in J} \sum_{t=L_j+1}^{T} \left(\left(T-t+1 \right) h_j \right) R_{j,t-L_j} \tag{5.41}$$

subject to

$$\sum_{\tau=1}^{t} R_{j,\tau-L_j} \geq \sum_{\tau=1}^{t} D_{j\tau} - I_{j0}, \quad \forall j \in J, \quad t = 1,\ldots,T \quad \left(\gamma_{jt} \right) \tag{5.42}$$

$$\sum_{j\in J(k)} a_{jk} R_{t-L_{jk}} \leq C_{kt}, \quad \forall k \in K, \quad t = 1,\ldots,T \quad \left(\sigma_{kt} \right) \tag{5.43}$$

$$R_{jt} \geq 0, \quad \forall j \in J, \quad t = L_j,\ldots,T-L_j \tag{5.44}$$

The Greek letters in parentheses denote the dual variables associated with each constraint set. The dual of this linear program is given by

$$\max \sum_{t=1}^{T} \sum_{j\in J} \left[\left(\sum_{\tau=1}^{t} D_{j\tau} - I_{j0} \right) \gamma_{jt} - \sum_{k\in K} C_{kt} \sigma_{kt} \right] \tag{5.45}$$

subject to

$$\begin{gathered} \sum_{\tau=t}^{T} \gamma_{j\tau} - \sum_{k\in K(j)} a_{jk} \sigma_{k,t-\left(L_j - L_{jk}\right)} \leq \left(T-t+1 \right) h_j, \quad \left(R_{j,t-L_j} \right) \\ \forall j \in J, \quad t = \left(L_j - L_{jk} \right) + 1,\ldots,T \end{gathered} \tag{5.46}$$

$$\gamma_{jt} \geq 0, \quad \forall j \in J, \quad t = 1,\ldots,T \tag{5.47}$$

$$\sigma_{kt} \geq 0, \quad \forall k \in K, \quad t = 1,\ldots,T \tag{5.48}$$

The primal variables corresponding to the dual constraints are shown next to each dual constraint set. While the primal problem chooses releases R_{jt} in each period t to minimize the cost of meeting demand under capacity constraints, the dual problem chooses prices γ_{it} and σ_{kt} to maximize revenue. γ_{jt} can be interpreted (subject to the mathematical caveats discussed by Rubin and Wagner (1990) and others) as the minimum amount the firm should charge an additional unit of demand for item j in period t. σ_{kt}, on the other hand, represents the maximum amount the firm should be willing to pay to acquire an additional unit of resource k in period t. The cost coefficients $(T - t + 1)h_j$ of the primal problem represent the contribution to the total cost of a unit of item j produced in period t, given by its incremental contribution to holding cost until the end of the planning horizon.

The first term in (5.45) represents the revenue from an additional unit of demand for item j in period t, which will increase the cumulative net demand $\sum_{\tau=1}^{t} D_{j\tau} - I_{j0}$ in each subsequent period until the end of the horizon. The second term in (5.45) represents the marginal cost of all resources required to process this additional unit of demand; recall that all demands must be met without backlogging. Hence the right-hand side of (5.46) represents the net marginal revenue (marginal revenue minus marginal resource costs) associated with an additional unit of demand for item j in period t. (5.46) ensures that the total marginal cost of the additional item cannot exceed its marginal net revenue. The complementary slackness property (5.40) implies that when there is positive slack in constraint (5.46) for some item j and period t at optimality, we will have $R_{jt} = 0$ in an optimal solution. Conversely, $R_{jt} > 0$ at optimality implies (5.46) is satisfied at equality.

Our primary interest in this discussion is the dual variables σ_{kt} associated with the primal capacity constraints (5.43). These dual variables represent the impact on the objective function of an additional unit of capacity at resource k in period t, which is of interest for several reasons. A high value of this dual variable indicates that limited capacity at this machine is significantly affecting the ability of the production unit to meet demand in a cost-effective manner, suggesting particular attention by management to improving its performance. It will also turn out, as we shall see in Chap. 7, that the clearing function formulations introduced in that chapter yield much more informative dual information than that obtained from this model, as we shall illustrate below.

Recall that a unit of product j that completes processing in period t will consume capacity on its k'th workcenter in period $t - L_j + L_{jk}$. Thus the output $X_{jt} = R_{j,t-L_j}$ of each item j in any period t is potentially constrained by at most $|K(j)|$ of the capacity constraints (5.43), each corresponding to a workcenter k in period $t - L_j + L_{jk}$. To ensure a non-degenerate optimal solution, we shall assume that for each item j at most one of these associated capacity constraints is satisfied at equality; this condition can be enforced if necessary by perturbing the right-hand side of the constraints by an arbitrarily small quantity. The specific workcenter k whose capacity constraint is binding in period $t - L_j + L_{jk}$ will be denoted by $k*(j,t)$, indicating that this workcenter limits the output of item j in period t. We will refer to resource $k*(j,t)$ as the limiting workcenter for item j in period t. Our assumption of non-degeneracy implies at most one limiting workcenter for each product j in each period t of the planning horizon. The limiting workcenter of an item j may be used concurrently by other items and may change from period to period, i.e., it is perfectly possible to have $k*(j,t) \neq k*(j,s)$ for $t \neq s$. Different items j may have different limiting resources in a given period.

As long as the production costs c_t are non-decreasing in the time period t, it is straightforward to show that an optimal solution to the primal will satisfy

$$\left(\sum_{j\in J(k)} a_{j,k^*(j,t)} R_{j,t-L_j+L_{j,k^*(j,t)}} - C_{k^*(j,t),t}\right) I_{j,t-1} = 0 \tag{5.49}$$

implying that the model will hold finished inventory against future demand in period $t - 1$ if and only if capacity at a resource *k*(j,t)* is fully utilized in period $t - L_j + L_{j,k^*(j,t)}$. Hence the model will hold finished inventory of product *j* in some period *t* only if the total demand for all items $\sum_{j\in J} D_{js}$ in some future period $s > t$ overloads the available capacity on its limiting resource *k*(j,s)* for period *s,* i.e.,

$$\sum_{j\in J} D_{js} > C_{k^*(j,s),s} \tag{5.50}$$

requiring the model to meet the demand in period *s* by building up finished inventory in periods prior to *s*. Periods *s* into which no finished inventory is carried in the optimal solution indicate that the optimal decisions for periods $s < t$ are independent of those for periods $s \geq t$. Hence an optimal solution to (5.41)–(5.44) will consist of one or more busy intervals, each consisting of $q \geq 0$ consecutive periods $S = \{s-q, s-q+1, \ldots, s\}$ with $I_{jq} > 0$ for some items $j \in J$ such that

$$\sum_{j\in J\left(k^*(j,s)\right)} a_{jk} R_{j,s-L_j+L_{j,k^*(j,s)}} = C_{k^*(j,s),s} \tag{5.51}$$

Since the limiting workcenter *k*(j,s)* has a binding capacity constraint in period *s* by definition, our assumption of a non-degenerate solution implies that all products $i \neq j$ requiring this resource in this period will have positive inventory in this busy interval, implying that $\gamma_{jt} = 0$ by the complementary slackness condition for constraints (5.42). Based on these observations, we will have dual prices $\sigma_{kt} > 0$ in periods $(s-q)-\left(L_j - L_{jk^*}\right)$ for all products *j* that use workcenter *k*(j,s)* in period *s*. We shall restrict our attention to periods in this interval where production activity is taking place, i.e., $X_{jt} = R_{j,t-L_j} > 0$. By complementary slackness, the dual constraints (5.46) will be tight in periods $s - q$ through *s*. Solving recursively from period $s + 1$ backwards in time to period *s*, Kefeli (2011) shows that

$$a_{j,k^*(j,t-1)} \sigma_{k^*(j,t-1),t-L_j+L_{j,}k^*(j,t-1)} - a_{j,k^*(j,t)} \sigma_{k^*(j,t),t-L_j+L_{j,}k^*(j,t)} = h_j \tag{5.52}$$

Our assumption of non-degeneracy implies that the output of each product *j* is limited by at most one resource *k* in each period, but there may be multiple workcenters with positive dual prices corresponding to different subsets of products. The limiting workcenter for a product *j* may also change from one period to the next, i.e., *k*(j,t − 1) ≠ k*(j,t).* When the same workcenter is limiting for item *j* in two consecutive periods $t - 1$ and *t*, (5.52) simplifies to

$$\left(\sigma_{k^*(j,t),(t-1)-L_j+L_{j,}k^*(j,t)} - \sigma_{k^*(j,t),t-L_j+L_{j,}k^*(j,t)}\right) = \frac{h_j}{a_{j,k^*(j,t)}} \tag{5.53}$$

Examining this expression shows that the absolute value of the dual price of capacity increases linearly with time over the busy interval, starting with a value of zero and increasing in absolute value by $h_j / a_{j,k^*(j,t)}$ in each period. This is intuitive; an additional unit of capacity at workcenter *k*(j,t)* in period $t - L_{j,k^*(j,t)}$ will allow $1/a_{j,k^*(j,t)}$ units held in inventory in period *s* to be produced in this period, reducing holding costs by $(t - s + 1)/a_{j,k^*(j,t)}$.

The following numerical example illustrates this structure of optimal solutions.

Example 5.1 To illustrate the structure of the optimal solution and the dual variables, consider a production unit with two products and four workcenters with the data given in Table 5.1. The unit finished goods holding costs are given by $h_1 = h_2 = 5$, and the overall lead times by $L_1 = L_2 = 5$. Initial inventory at the end of period 0 is $I_{10} = 20$ units for Product 1 and $I_{20} = 25$ units for Product 2. Both products require processing on all four resources in increasing order of machine number. The demand for each product in each period is given in Table 5.2.

Table 5.1 Parameter values for Example 5.1

	Item	Machine 1	Machine 2	Machine 3	Machine 4
Production cost	1	1	1	1	2
	2	2	2	2	2
Processing time	1	3	3	2	4
	2	3	4	4	4
Lead time L_{jk}	1	0	1	2	4
	2	0	1	2	4
Capacity C_{kt}		100	100	18	20

Table 5.2 Demand data for Example 5.1

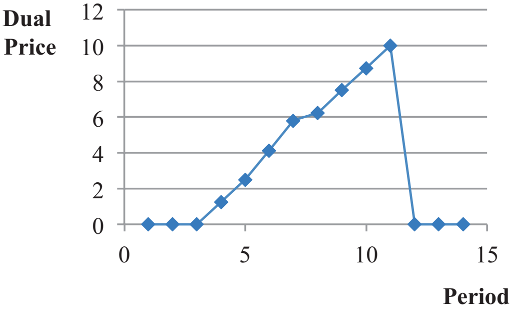
Period	Item 1 demand	Item 2 demand
1	0	0
2	0	0
3	5	0
4	4	0
5	4	2
6	5	4
7	5	5
8	5	5
9	6	3
10	7	4
11	6	3
12	6	3
13	0	6
14	0	5
15	0	0

Solving the primal model (5.41)–(5.44) yields the optimal solution given in Table 5.3, with optimal objective function value 2636.25. Machine 3 is the limiting resource for Product 1 in periods 4 through 8, and for Product 2 in periods 9 and 10. The dual prices σ_{kt} associated with this machine in the optimal solution are plotted in Fig. 5.4; only Machine 3 has nonzero dual prices. The relation (5.53) can be clearly observed, with the dual price increasing linearly until the capacity loading falls below resource capacity. Note that although Machine 4 has utilization of 0.9 in periods 12 and 13, and would thus be expected to have high cycle time and WIP, its dual price remains at zero since the capacity constraint is not binding.

Table 5.3 Optimal solution for Example 5.1

	Releases, R_j		Resource loading				Ending inventory, I_{jt}	
Period	Item 1	Item 2	Machine 1	Machine 2	Machine 3	Machine 4	Item 1	Item 2
−3								
−2								
−1								
0							20	25
1	0	3.75	11.25	0	0	0	20	25
2	3	2.25	15.75	15	0	0	20	25
3	6	0	18	18	15	0	15	25
4	6	0	18	18	18	0	11	25
5	6	0	18	18	18	15	7	23
6	6	0	18	18	18	15	2	22.75
7	6	0	18	18	18	12	0	20
8	0	4.5	13.5	18	18	12	1	15
9	0	4.5	13.5	18	18	12	1	12
10	0	0	0	18	18	12	0	8
11	0	0	0	0	18	12	0	5
12	0	0	0	0	0	18	0	2
13	0	0	0	0	0	18	0	0.5
14	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0

Fig. 5.4 Dual variables associated with Machine 3 in optimal solution to Example 5.1



5.4.1 Insights from the Dual

Our analysis of the dual solution indicates a number of drawbacks of the formulation (5.41)–(5.44), particularly its representation of workcenter behavior. By complementary slackness, the dual variables σ_{kt} associated with the capacity constraints (5.43) will only take nonzero values if the associated primal constraint is binding at optimality, implying that the workcenter's capacity is fully utilized. Since σ_{kt} represents the maximum amount the firm should be willing to pay for an additional unit of output from workcenter k in period t, this implies that no improvement in the optimal objective function value can be obtained from additional output at a workcenter unless its capacity is fully utilized. However, as discussed in Chap. 2, queueing models suggest qualitative changes in the behavior of a capacitated workcenter at utilization levels well below 1; more precisely, they show a nonlinear increasing relation proportional to $1/(1-u)$ between cycle time and utilization (Hopp and Spearman 2008), implying that additional capacity at the workcenter might improve system performance even though currently capacity is not fully utilized. Likewise, improving the performance of a workcenter such that it can generate more output for the same average WIP level, shifting the curves in Fig. 2.2 to the left, should allow reduced cycle time and hence reduced costs, which the current model is unable to capture. Note, however, that this does not necessarily imply that adding capacity would be economically beneficial, especially if capacity can only be added in discrete increments.

A second drawback of this model can be observed directly in (5.53): the dual price of a resource in a period is independent of events at other resources as long as the limiting resources do not change. This again contradicts insights from queueing models (Hopp and Spearman 2008), which show that the behavior of downstream resources is affected by the utilization of upstream ones. Consider two resources operating in series where work flows from workcenter 1 to workcenter 2. Per Hopp and Spearman (2008) Chap. 8, the squared coefficient of variation (SCV) of the interarrival times at workcenter 2 is given by the SCV of the departure process from workcenter 1, which, in turn, can be approximated by

$$c_{\mathrm{d}}^2 = u^2 c_{\mathrm{e}}^2 + \left(1-u^2\right) c_{\mathrm{a}}^2 \tag{5.54}$$

where u denotes the average utilization of workcenter 1, c_{e}^2 the SCV of the effective processing time distribution at workcenter 1, and c_{a}^2 the SCV of the external arrival process to workcenter 1. This relation suggests that the dual price of capacity at workcenter 2 ought to be influenced by decisions at workcenter 1; under most conditions, unless c_{e} is small relative to c_{a}, adding capacity to workcenter 1 will reduce u, reducing the average cycle time at workcenter 2 which ought to improve overall performance, or at least leave it no worse.

This analysis of the dual prices of capacity suggests that the use of fixed lead times can model the behavior of production resources subject to queueing behavior to at best a limited degree. The largest discrepancies are to be expected when resource utilization levels vary significantly over time, causing the fixed lead times to over- and/or underestimate actual cycle times; and when multiple resources have high utilization levels close to, but not quite equal to, 1, such that small changes in utilization lead to large changes in cycle times.

5.5 Fractional Lead Times

Our discussion of fixed lead times up to this point has assumed lead times expressed as integer multiples of the planning period length Δ, recalling that period t ends at time $t\Delta$. Assuming that all release and demand rates are uniform over each planning period, Hackman and Leachman (1989b) have shown that non-integer fixed lead times can be incorporated easily. We first illustrate the basic idea with a single-product single-workcenter model and then discuss generalizations.

Any fractional fixed lead time L can be decomposed into integer and fractional parts as $L = \lfloor L \rfloor + \phi$, where $\lfloor L \rfloor$ denotes the largest integer less than or equal to L and $\phi = L - \lfloor L \rfloor$ the fractional part of the lead time. We assume L remains constant in all planning periods; the case where lead times can vary over time is addressed in the next chapter. Under uniform release and demand rates over the planning period, if R_t units of a product are released during this period, the material will enter the production unit at a rate of R_t/Δ units per unit time. The material flow through the workcenter can then be represented as in Fig. 5.5. The upper timeline represents the progression of releases into the production unit over time and the lower timeline the entry of this material into finished inventory. The amount of material becoming available to meet demand in period t is given by

$$Y_t = \phi R_{t-\lfloor L \rfloor -1} + (1-\phi) R_{t-\lfloor L \rfloor} \tag{5.55}$$

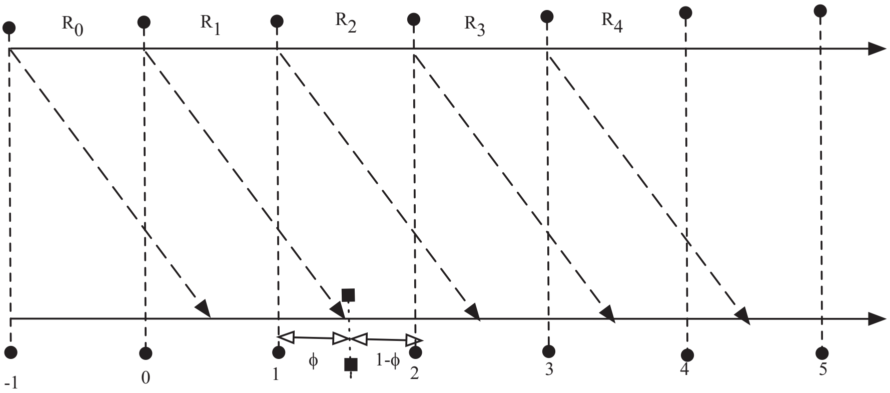


Fig. 5.5 Fractional lead times

Table 5.4 Data for fractional lead time example

Period	Releases	Outputs	Demand	Ending inventory
−1	50		–	–
0	80		–	0
1	120	65	50	15
2	150	100	110	5
3	150	135	140	0
4	160	150	150	0
5	150	155	155	0
6	80	155	150	5
7	25	115	100	20
8	0	52.5	60	12.5
9	0	12.5	25	0

and the material balance constraint analogous to (5.16) takes the form

$$I_t = I_{t-1} + \phi R_{t-\lfloor L \rfloor -1} + (1-\phi) R_{t-\lfloor L \rfloor} - D_t, \quad t = 1,\ldots,T \tag{5.56}$$

where I_t denotes the amount of finished goods inventory at the end of period t.

However, let us take a closer look at the implications of this formulation. Recall that we assume constant release and production rates throughout each planning period. Now consider the data given in Table 5.4, under a fixed lead time of $L = 1.5$ periods and $R_t = 0$ for $t < -1$.

The output (production) in each period t is computed assuming that releases R_t and demands D_t are uniformly distributed across their associated planning periods as in (5.56). The ending inventory is computed using the inventory balance equation (5.56) at the end of each period. The reader should verify these calculations to confirm that inventory levels are nonnegative at the end of all planning periods.

However, all is not as it seems. Although the release rate over each planning period is constant, the output rate, which defines the rate of inflow into the inventory, is not. Due to the fractional nature of the lead time, material released at the start of period t emerges as output in the middle of period $t + 1$, as illustrated in Fig. 5.6, where each period is divided into two subintervals of length ϕ and $1 - \phi$, in this case both equal to 0.5 periods. In periods 1, 2, and 3, the output rate of the production resource during the first subinterval of the period is different from that in the second subinterval.

Table 5.5 recalculates Table 5.4 at each half-period. As the reader can (and should!) verify, changes in output rates within the planning periods result in negative inventory levels at some of these intermediate points.

As pointed out by Hackman and Leachman (1989b), there are two possible solutions to this problem. The most obvious, especially in the very structured example we have used here, is to reduce the size of the planning periods such that rate changes within planning periods are no longer possible, and enforce material balance and capacity constraints at the boundaries of each of these subintervals.

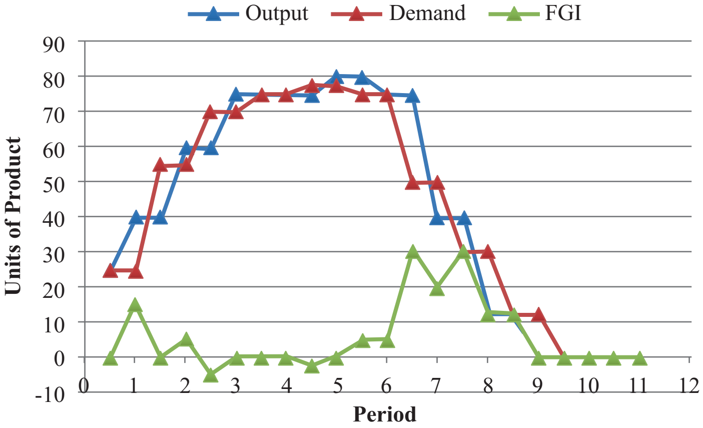


Fig. 5.6 Output of production unit with fractional lead times per Table 5.5

Table 5.5 Effect of fractional lead times at interior points of planning periods

Period	Output	Demand	FGI
0.5	25	25	0
1	40	25	15
1.5	40	55	0
2	60	55	5
2.5	60	70	-5
3	75	70	0
3.5	75	75	0
4	75	75	0
4.5	75	77.5	−2.5
5	80	77.5	0
5.5	80	75	5
6	75	75	5
6.5	75	50	30
7	40	50	20
7.5	40	30	30
8	12.5	30	12.5
8.5	12.5	12.5	12.5
9	0	12.5	0
9.5	0	0	0
10	0	0	0
10.5	0	0	0
11	0	0	0

To ensure consistency, the length of the subintervals must be equal to the least common divisor of the fractional lead times. This will significantly increase the size of the formulation, since both the number of constraints and the number of decision variables depend on the number of periods. It is also impractical in the presence of the time-varying lead times discussed in Chap. 6, where the lead time associated with each period may have a different fractional part. In this case a period length equal to the greatest common divisor of all lead times must be used, which may result in a much larger model than necessary.

Hackman and Leachman (1989b) propose a much simpler solution to this difficulty by noting that it is only necessary to write additional constraints at points in time where output or release rates may change. This includes the boundaries of the original planning periods and intermediate points where a fractional lead time causes a change in output (and hence the rate of inflow into finished inventory) or the amount of material requiring capacity at a particular resource. Under the time-stationary lead times assumed in this chapter, each planning period will have at most one intermediate point for which additional constraints for a given product need to be written.

Although we have focused on the overall lead time L_j of the production unit for a particular item j, the same issues arise with respect to the capacity constraints for each workcenter k and their associated lead times L_k. In this case the changes in release rate within a planning period may result in capacity constraints being violated at interior points of the period (Hackman and Leachman 1989b). To see this, consider the situation illustrated in Fig. 5.7 where we have two items whose respective lead times are $L_{1k} = 1.5$ and $L_{2k} = 1.75$ periods. The upper time line shows the releases of each item and the lower the arrival of each item at the resource under consideration. Recalling our convention that period t ends at time t, the rate of material arriving at the resource k during period t can change at three potential points in time: $t - 1$, $t + \phi_{1k}$, and $t + \phi_{2k}$, where $\phi_{jk} = L_{jk} - \lfloor L_{jk} \rfloor$, requiring the capacity constraints

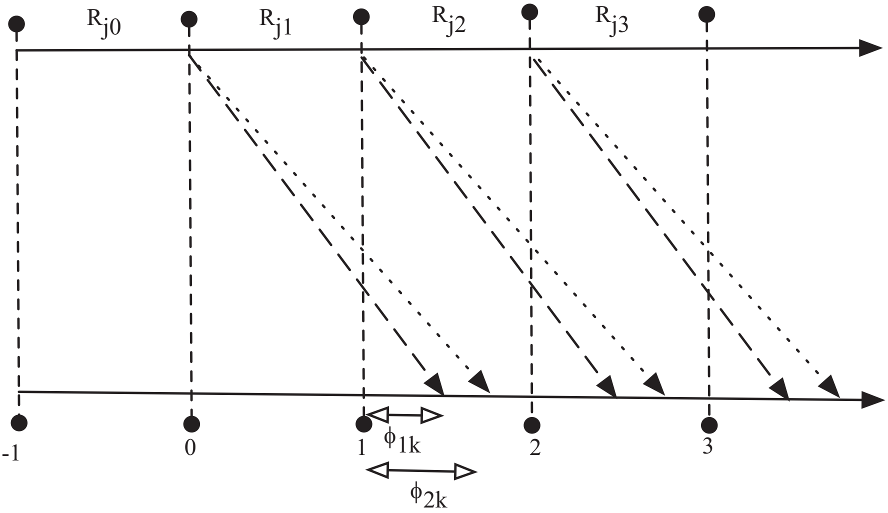


Fig. 5.7 Impact of multiple products with fractional lead times on capacity constraints

$$\begin{aligned}\phi_{1k}\left(a_{1k}R_{t-L_{1k}-1}+a_{2k}R_{t-L_{2k}-1}\right)&\le\phi_{1k}C_{kt}\\ \left(\phi_{2k}-\phi_{1k}\right)\left(a_{1k}R_{t-L_{1k}}+a_{2k}R_{t-L_{2k}-1}\right)&\le\left(\phi_{2k}-\phi_{1k}\right)C_{kt}\\ \left(1-\phi_{2k}\right)\left(a_{1k}R_{t-L_{1k}}+a_{2k}R_{t-L_{2k}}\right)&\le\left(1-\phi_{2k}\right)C_{kt}\end{aligned}\tag{5.57}$$

This approach results in a large number of additional capacity constraints, especially in environments such as semiconductor wafer fabrication where a given resource may be used by tens of different unit operations. Leachman (2001) points out that the presence of many items j with slightly different fractional components ϕ_{jk} is likely to yield a roughly uniform distribution of workload over the planning interval, allowing approximate capacity constraints of the form

$$\sum_{j\in J(k)}a_{jk}\left[\phi_{jk}R_{t-L_{jk}-1}+\left(1-\phi_{jk}\right)R_{t-L_{jk}}\right]\le C_{kt},\quad t=1,\ldots T;\quad k=1,\ldots,K\tag{5.58}$$

to be used without inducing excessive error. Note that (5.58) simply adds up the total amount of each product loading the resource within the planning period, without considering the specific timing of the loading within the period. The basic operation of these constraints is the same as that for material flow discussed above and can be illustrated in the following example.

Example 5.2 Consider a single resource and three products with fixed fractional lead times $L_1 = 1.3$, $L_2 = 1.5$, and $L_3 = 1.75$ that remain constant over a planning horizon consisting of $T = 12$ periods. Thus we have $\phi_1 = 1.3 - \lfloor 1.3 \rfloor = 0.3$, $\phi_2 = 0.5$, and $\phi_3 = 0.75$ by the same logic. Following Fig. 5.7, the intervals within which the capacity loading from each product will remain constant, assuming constant release rates over each planning period, are calculated in Table 5.6.

Capacity loading of the resource remains constant over each interval with the given start and end points. Due to the fractional lead times, the rate of capacity

Table 5.6 Uniform loading intervals for Example 5.2

Prod. 1	Start	0.3	1	1.3	2	2.3	3	3.3	4	4.3	5	5.3	6	6.3
	End	1	1.3	2	2.3	3	3.3	4	4.3	5	5.3	6	6.3	7
Prod. 2	Start	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5
	End	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7
Prod. 3	Start	0.75	1	1.75	2	2.75	3	3.75	4	4.75	5	5.75	6	6.75
	End	1	1.75	2	2.75	3	3.75	4	4.75	5	5.75	6	6.75	7
Prod. 1	Start	7	7.3	8	8.3	9	9.3	10	10.3	11	11.3	12	12.3	
	End	7.3	8	8.3	9	9.3	10	10.3	11	11.3	12	12.3	13.3	
Prod. 2	Start	7	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12	12.5	
	End	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12	12.5	13.5	
Prod. 3	Start	7	7.75	8	8.75	9	9.75	10	10.8	11	11.8	12	12.8	
	End	7.75	8	8.75	9	9.75	10	10.8	11	11.8	12	12.8	13.8	

loading can change at the start of any of these intervals for any product; hence it is necessary to write capacity constraints for each subinterval arising from the intersections of consecutive loading intervals for the individual products. Thus constraints similar to (5.57) need to be written for the intervals (0.3, 0.5), (0.5, 0.75), (0.75, 1), (1, 1.3), (1.3, 1.5), and so on, resulting in a total of 53 intervals that need to be considered explicitly. The large number of constraints required by this approach is immediately evident.

Assuming that each unit of each product requires a single unit of capacity for one planning period to be completed, we now calculate the capacity loading, in terms of units of capacity, for each interval assuming the releases in Table 5.7.

Figure 5.8 plots the total loading of the resource by all three products for the release schedule shown in Table 5.7. The interval load plot shows the load, in terms of the number of parallel machines that would be required to process all the work available in the interval, for each of the 53 subintervals over which load remains constant, while the period load plots the total load within each planning period using (5.58). Discrepancies between the two plots arise where one would expect, in regions where the releases, and hence the loading of the resources, is changing, which in the example are at the start and end of the planning horizon and between periods 5 and 6, where the releases of Product 2 are temporarily interrupted. As the number of products and the number of different ϕ_j values increase, and especially if the ϕ_j values are distributed somewhat uniformly between 0 and 1, the error induced

Table 5.7 Release schedule for Example 5.2

Period	0	1	2	3	4	5	6	7	8	9	10	11	12
Prod. 1	1	1	1	1	1	1	1	1	1	1	1	1	1
Prod. 2	2	2	2	2	1	0	0	0	2	2	2	2	2
Prod. 3	0	0	0	0	0	0	1	1	1	1	0	0	0

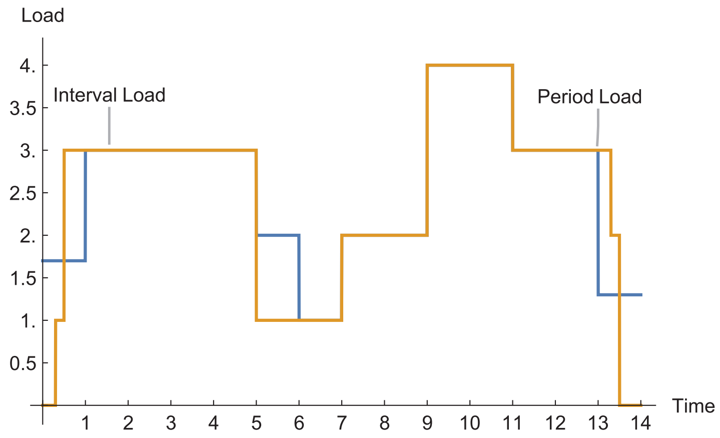


Fig. 5.8 Capacity loading with fractional lead times

by using (5.58) is likely to be considerably smaller than that arising from other sources, such as errors in demand forecasts.

Kacar et al. (2016) compare the performance of planning models using integer and fractional lead times on a data set representing a large semiconductor wafer fabrication facility and find that incorporating fractional lead times for both finished inventory balance and capacity constraints yields markedly superior performance than including it for either one alone. The fractional lead time model including the aggregate capacity constraints (5.58) and the finished inventory balance constraints (5.56) yielded significantly better performance than a model using integer lead times, and comparable performance to the much larger model using clearing functions described in Chap. 7.

Given the magnitude of the performance improvement and the capabilities of today's commercial LP solvers, we see no reason not to use fractional lead times if they appear to be called for. The most likely case where fractional lead times will be beneficial is when the cycle times of the production system and individual resources span multiple planning periods, and the fractional parts of the lead times are substantial relative to the length of the planning period.

5.6 Input-Output Models: An Alternative View of Fixed Lead Times

Our discussion so far has assumed that the entire quantity R_{jt} of product j released into the production unit in period t (the production orders released in period t) moves through the production unit as a single entity, such that all items released in that period consume capacity and enter finished inventory together. Given the assumption of releases taking place at a constant rate over the planning periods, each unit of product j will be processed at resource k L_{jk} time units after its release. Hence these lead times represent the time elapsing between the release of the material to the first resource on its routing and its consuming capacity on resource k. Under integer lead times, this implies that all materials released in period t will consume capacity at resource k in period $t + L_{jk,}$, i.e., at the end of the specified lead time; the case of fractional lead times is a simple extension of this idea as discussed in the previous section.

An alternative view of fixed lead times allows a production order to consume capacity on a workcenter anywhere within the time it is expected to spend at the workcenter estimated by (5.11). This requires defining new lead time parameters L'_{jk} representing the arrival times of the orders at the workcenters, i.e., the earliest possible time after its release that processing of the material at the k'th resource on its routing can start. Note that the lead times L_{jk} we have used in the previous sections represent a different quantity, the time elapsing between release and capacity consumption. Thus a production order of product j released in period t can consume capacity on the k'th workcenter in its routing anywhere in the time interval $[t + L'_{jk},$

$t + L'_{j,k+1}-1$], instead of in period $t + L_{jk}$. Management of the production unit may elect to process portions of a production order in several, not necessarily consecutive, periods, while still ensuring completion of the order within its planned lead time L_j. This timing flexibility reflects the possibility of production smoothing within the lead times through scheduling decisions, whereas in the models in the previous sections production quantities are entirely determined by the releases. Models of this kind have been proposed in several different contexts. Pürgstaller and Missbauer (2012) note that the Input-Output Control approach of Wight (1970) implies a model of this form, although the model is not explicitly stated. We have also shown in Chap. 4 that a similar model is implicit in the LUMS order release mechanism for make-to-order production (Hendry et al. 2013). The structure of these models is also related to a much older formulation by Bowman (1956). Spitter et al. (2005) and de Kok and Fransoo (2003) consider a production unit with a single bottleneck workcenter that may consist of a number of parallel machines. In these latter papers, the primary purpose of the model is for supply chain coordination rather than detailed release planning, so they do not directly accommodate modeling of production flows across multiple resources within a production unit. The formulation given below extends these models to incorporate such production flows.

Since the release quantities R_{jt} no longer define the capacity loading of resources in a unique manner, we define additional decision variables Z^k_{jts} specifying the amount of product j released in period t that consumes capacity on workcenter k in period s. To ensure that the workcenters in the routing are visited in the correct sequence, we must define these variables to ensure that processing on workcenter k can only take place in the correct time interval such that $t + L'_{jk} \le s \le t + L'_{j,k+1} - 1$ and $1 \le t \le T - L_j$.

Since all materials entering the system must be processed on every workcenter (neglecting details such as scrap or yield losses), we have

$$R_{jt} = \sum_{k\in K} \sum_{s=t+L'_{jk}}^{t+L'_{j,k+1}-1} Z^k_{jts}, \forall j \in J, s = 1,\ldots,T \tag{5.59}$$

Since the processing of a given production order may now be distributed over several periods, all materials associated with the production order released in period t need not necessarily enter finished inventory together. If the production order can enter finished inventory only after the planned lead time has elapsed, irrespective of the actual time(s) the material is processed, the finished inventory balance equations will take the form

$$I_{jt} = I_{j,t-1} + R_{j,t-L_j} - D_{jt}, \quad t = 1,\ldots,T, \quad j \in J \tag{5.60}$$

If, however, material can enter finished inventory as it completes its processing, without having to wait for the remainder of the order, the finished inventory balance equation will be

$$I_{jt} = I_{j,t-1} + \sum_{s=t-L'_{j,K(j)}}^{t} Z^{K(j)}_{jst} - D_{jt}, \quad \forall j \in J, \quad t = 1,\ldots,T \tag{5.61}$$

where *K(j)* denotes the last resource in the process routing of item *j*. Since (5.60) is more consistent with the intent of a planned lead time to ensure availability of the material after the planned lead time with high probability while leaving internal resource allocation decisions to the local management, we shall adopt this assumption from now on. The R_{jt} variables can, of course, be eliminated using (5.59) to reduce the number of variables when solving the model.

The capacity constraints for each workcenter *k* will now take the form

$$\sum_{j \in J} \sum_{s=t-L'_{jk}}^{t} a_{jk} Z_{jst}^{k} \le C_{kt}, \quad t = 1,\ldots,T-L_j, \quad \forall k \in K \tag{5.62}$$

where the summation on the left hand side represents the total amount of work allocated to workcenter *k* in period *t*. Hence while it is possible to incorporate time-dependent production costs at the different workcenters, if costs are time-stationary there is no need to do so due to the no backlogging assumption. The complete formulation can now be written as

$$\min \sum_{j \in J} \sum_{t=1}^{T-L_j} h_j I_{jt} \tag{5.63}$$

subject to (5.59), (5.60) or (5.61), (5.62) depending on assumptions, and

$$R_{jt} \ge 0, \forall j \in J, t = 1,\ldots,T \tag{5.64}$$

$$Z_{jst}^{k} \ge 0, \forall j \in J, \forall k \in K, t = 1,\ldots,T-L_j, s = t-L_{jk},\ldots,t \tag{5.65}$$

Since imposing the additional constraint that

$$Z_{jst}^{k} = \begin{cases} R_{jt}, \text{for } s = t - L_{jk} \\ 0, \text{otherwise} \end{cases} \tag{5.66}$$

with L_{jk} denoting the pre-specified lead time in the interval $[L'_{jk}; L'_{j,k+1} - 1]$ recovers formulation (5.26)–(5.29), (5.63)–(5.65) is a relaxation of the former in the sense that any feasible solution to (5.26)–(5.29) is feasible for (5.63)–(5.65), but not vice versa. As with formulation (5.26)–(5.29), (5.63)–(5.65) can be rewritten to eliminate the I_{jt} variables giving a model analogous to (5.41)–(5.44).

The formulation until this point has ignored WIP costs. Their inclusion requires some additional thought. If material released at *t* and processed at workcenter *k* cannot move to the next workcenter in its routing until time $t + L'_{j,k+1}$, two different types of WIP may exist at a workcenter: material that has been processed and is waiting to move to the next stage and material that has not yet been processed. If the value of the WIP depends on the timing of production that results from the mode, i.e., earlier production within the lead time means higher WIP holding costs, this can be accounted for by decomposing the WIP at the workcenters into WIP before and WIP after processing and assigning different WIP holding costs to each component.

Denoting W^{b}_{jkt} and W^{a}_{jkt} the WIP of product *j* at workcenter *k* at the end of period *t* before and after processing, respectively, the WIP balance equations are

$$W^{\mathrm{b}}_{jkt} = W^{\mathrm{b}}_{j,k,t-1} + R_{j,t-L'_{jk}} - \sum_{s=t-L'_{jk}}^{t} Z^{k}_{jst}, \quad \forall j \in J; \quad t = 1,\ldots,T; \quad k \in K \tag{5.67}$$

$$W^{\mathrm{a}}_{jkt} = W^{\mathrm{a}}_{j,k,t-1} + \sum_{s=t-L'_{jk}}^{t} Z^{k}_{jst} - R_{j,t-L'_{j,k+1}}, \quad \forall j \in J; \quad t = 1,\ldots,T; \quad k \in K \tag{5.68}$$

where in (5.68) $L'_{j,|K|+1} = L_j$. The complete model formulation is given by Pürgstaller and Missbauer (2012). We do not consider this extension in the following example, but note it as an illustration of an issue that the more flexible treatment of lead times may raise.

Example 5.3 We implement the model (5.63)–(5.65) with the finished inventory balance constraints (5.60) on the problem instance solved in Example 5.1, where we set the values of the L'_{jk} to the L_{jk} values in that example. An optimal solution with objective function value 2505 is obtained as shown in Table 5.8.

The principal difference, as one would expect, lies in the distribution of the capacity loading on Machine 3. Since this machine has a local lead time of $L'_{j4} - L'_{j3} = 4 - 2 = 2$ periods for both products, it is able to allocate capacity across two different periods to releases made in a single period, unlike the previous model where all releases from a given period *t* will load a resource in the single period $t + L_{jk}$ (assuming integer lead times). This difference is seen in Table 5.9 that shows

Table 5.8 Optimal solution for Example 5.3

	Releases		Capacity loading				Ending inventory	
Period	Item 1	Item 2	Machine 1	Machine 2	Machine 3	Machine 4	11	12
0	0	0	0	0	0	0	20	25
1	3	1.5	13.5	0	0	0	20	25
2	6	0	18	15	0	0	15	25
3	6	0	18	18	15	0	11	25
4	6	0	18	18	18	0	7	23
5	6	1.5	22.5	18	18	12	2	19
6	6	2	24	24	18	12	0	15.5
7	0	5	15	26	18	12	1	10.5
8	0	5	15	20	18	12	1	7.5
9	0	0	0	20	18	18	0	3.5
10	0	0	0	0	18	20	0	2
11	0	0	0	0	18	20	0	1
12	0	0	0	0	0	20	0	0
13	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0

Table 5.9 Z^3_{jst} values for Machine 3 in Example 5.3

	Release	Loading period														
	Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Item 1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2		0	0	3	0	0	0	0	0	0	0	0	0	0	0
	3			0	0	6	0	0	0	0	0	0	0	0	0	0
	4				0	0	6	0	0	0	0	0	0	0	0	0
	5					0	0	6	0	0	0	0	0	0	0	0
	6						0	0	6	0	0	0	0	0	0	0
	7							0	0	4	2	0	0	0	0	0
	8								0	0	0	0	0	0	0	0
	9									0	0	0	0	0	0	0
	10										0	0	0	0	0	0
	11											0	0	0	0	0
	12												0	0	0	0
	13													0	0	0
	14														0	0
	15															0
Item 2		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2		0	0	1.5	0	0	0	0	0	0	0	0	0	0	0
	3			0	0	0	0	0	0	0	0	0	0	0	0	0
	4				0	0	0	0	0	0	0	0	0	0	0	0
	5					0	0	0	0	0	0	0	0	0	0	0
	6						0	0	0	1.5	0	0	0	0	0	0
	7							0	0	0	2	0	0	0	0	0
	8								0	0	1	4	0	0	0	0
	9									0	0	0.5	4.5	0	0	0
	10										0	0	0	0	0	0
	11											0	0	0	0	0
	12												0	0	0	0
	13													0	0	0
	14														0	0
	15															0

the values of the Z^3_{jst} variables for Machine 3. Releases of Product 1 made in period 7 are processed in periods 9 and 10; releases of Product 2 in period 8 are processed in periods 10 and 11, and those from period 9 in periods 11 and 12. Thus in period 11, releases of Product 2 from two different, but consecutive, periods are being processed.

The dual prices associated with this optimal solution are plotted in Fig. 5.9. Note that now both Machines 3 and 4 have binding capacity constraints and hence positive absolute dual prices, in the later periods of the planning horizon. While the

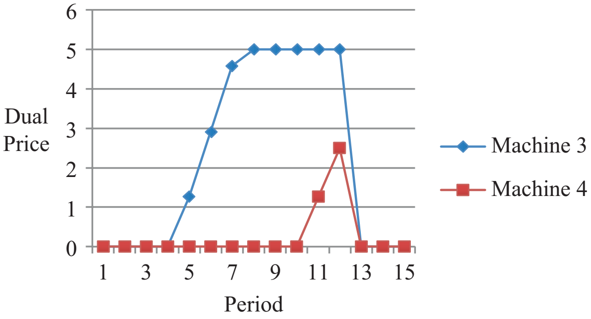


Fig. 5.9 Dual variables for Example 5.3

formulation in Example 5.1 results in positive dual prices for Machine 3 over the same time interval as in this case, Machine 4 never achieves a positive dual price under the previous formulation.

5.7 A Caveat: Lot-Sizing Issues

The models described so far yield the release quantities per product and period R_{jt} as the essential result that is used by the planning level. Executing these decisions in a straightforward manner means releasing production orders of size R_{jt} in the respective periods. However, if the sizes of the production orders are fixed, as is the case when production orders are released to the shop floor by an MRP system that specifies standard lot sizes, the R_{jt} quantities should be viewed as release "budgets" that are filled or consumed by the orders. This is also the case if the model is formulated for aggregate products or product families j with similar routing and resource requirements for the products within one family. Even if the demand D_{jt} is derived from the order sizes, the release quantities need not be a sum of the order sizes due to the capacity constraints and the real-valued R_{jt}. In this case the orders to release within the quantities R_{jt} must be determined by a separate planning step. One obvious possibility is to release the orders of product j in period t in the sequence of increasing due date until the cumulative actual release quantity reaches its planned value, that is

$$\sum_{\tau=1}^{t} Actual\ release\ quantity_{j\tau} \le \sum_{\tau=1}^{t} R_{j\tau},\ \forall j \in J,\ t = 1,\ldots,T \qquad (5.69)$$

perhaps with the possibility to exceed the cumulative planned releases by the last order as applied in Load-Oriented Order Release discussed in Sect. 4.2.2.

Alternatively, the release model can be formulated at the level of production orders $p = 1, \ldots, P_j$ with order size, due date, and capacity requirements (setup and processing time) given for each order. Without loss of generality, we assume that the orders are indexed in the order of increasing due dates. The model then determines the period in which each order will be released. The decision variables are

$$\delta_{jpt} = \begin{cases} 1, \text{ if order } p \text{ of product } j \text{ is released in period } t \\ 0, \text{ otherwise} \end{cases}.$$

The release periods are subject to the constraints

$$\sum_{t=1}^{T} \delta_{jpt} = 1, \quad \forall j; \quad \forall p = 1,\ldots,P_j \tag{5.70}$$

$$\sum_{\tau=1}^{t} \delta_{jp\tau} \geq \sum_{\tau=1}^{t} \delta_{j,p+1,\tau}, \quad \forall j,t; \quad \forall p = 1,\ldots,P_j - 1 \tag{5.71}$$

where (5.70) ensures that each order is released exactly once and (5.71) maintains the correct release sequence of the orders. The release quantities can be obtained by

$$R_{jt} = \sum_{p=1}^{P_j} Q_{jp} \delta_{jpt}, \quad \forall j,t \tag{5.72}$$

where Q_{jp} denotes the order size of order p of product j.

This modeling technique can be applied to both the conventional fixed lead time model (5.26)–(5.29) and for the alternative model with variable timing of production described in this section, which is described in Missbauer (2014). At the present time, there is no experience with the solvability of the resulting MILP model for real-life problems. Heuristics, e.g., decomposing by product and coordinating the resulting subproblems by Lagrangian techniques or by column generation, are an obvious possibility.

5.8 Summary and Conclusions

In this chapter, we have examined the structure of production planning models based on fixed, exogenous lead times that remain constant over the planning horizon. This constitutes the most prevalent mechanism for representing cycle times in both the research literature and industrial practice. We have shown that different models are possible depending on what assumptions are made on the timing of different events, such as when capacity is consumed on specific resources relative to the release time. We have also shown that models with fixed positioning of production within the lead time treat WIP in a rather restrictive manner, assuming WIP cannot accumulate and only a portion of the total WIP in the system is available to be processed by a resource in a given period.

We have also illustrated several limitations of these models relative to the behavior of production resources discussed in Chap. 2. Queueing models show that average cycle time is nonlinear in the average resource utilization, which is directly determined by the work release decisions made by planning models. However, fixed exogenous lead times ignore this relationship, assuming that as long as all capacity constraints are satisfied changes in cycle time due to workload will be negligible. Queueing models also suggest that cycle times begin to degrade well before utilization reaches 1, suggesting there may be benefit to additional capacity at resources whose utilization is below 1. However, our analysis of the dual prices of capacity shows that until a resource is fully utilized, dual prices will be zero, suggesting no benefit from additional resources.

The limited research examining the benefits of more sophisticated models with workload-dependent lead times (Kacar et al. 2012, 2013, 2016) suggests that as long as the average resource utilization remains relatively constant, fixed lead time models with appropriately chosen values of the lead times yield performance very similar to that of much more complex models with workload-dependent lead times. The use of fractional lead times yields a significant improvement over integer lead times, at little additional cost in model complexity. However, when resource utilization and product mix vary significantly over time, the performance of fixed lead time models begins to deteriorate. For this reason, as well as to address the theoretical drawbacks of fixed lead time models discussed above, it is of interest to explore planning models capable of recognizing the nonlinear relation between utilization and cycle time. Put another way, fixed lead time models optimize over releases only; queueing results suggest that jointly optimizing releases and lead times may yield better results. We now explore these more advanced models in the next chapters.

References

Baker KR (1993) Requirements planning. In: Graves SC, Kan AHGR, Zipkin PH (eds) Handbooks in operations research and management science. Logistics of production and inventory, vol 3. Elsevier Science, Amsterdam, pp 571–627

Bazaraa MS, Jarvis J, Sherali HD (2004) Linear programming and network flows. Wiley, New York

Bertsimas D, Tsitsiklis JN (1997) Introduction to linear optimization. Scientific, Athena

Billington PJ, Mcclain JO, Thomas JL (1983) Mathematical programming approaches to capacity-constrained MRP Systems: review, formulation and problem reduction. Manag Sci 29:1126–1141

Bowman EB (1956) Production scheduling by the transportation method of linear programming. Oper Res 4(1):100–103

de Kok AG, Fransoo JC (2003) Planning supply chain operations: definition and comparison of planning concepts. In: de Kok AG, Graves SC (eds) Handbooks in operations research and management science. Supply chain management: design, coordination and operation, vol 11. Elsevier, Amsterdam, pp 597–675

Hackman S (1990) An axiomatic framework of dynamic production. J Prod Anal 1:309–324

Hackman S (2008) Production economics. Springer, Berlin

Hackman S, Leachman RC (1989a) An aggregate model of project oriented production. IEEE Trans Syst Man Cybern 19(2):220–231

Hackman ST, Leachman RC (1989b) A general framework for modeling production. Manag Sci 35(4):478–495

Hanssmann F, Hess SW (1960) A linear programming approach to production and employment scheduling. Manag Technol 1(1):46–51

Hendry L, Huang Y, Stevenson M (2013) Workload control: successful implementation taking a contingency-based view of production planning and control. Int J Oper Prod Manag 33(1):69–103

Holt CC, Modigliani F, Simon HA (1955) A linear decision rule for production and employment scheduling. Manag Sci 2(1):1–30

Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston

Jacobs FR, Berry WL, Whybark DC, Vollmann TE (2011) Manufacturing planning and control for supply chain management. McGraw-Hill Irwin, New York

Jansen B, de Jong JJ, Roos C, Terlaky T (1997) Sensitivity analysis in linear programming: just be careful! Eur J Oper Res 101(1997):15–28

Johnson LA, Montgomery DC (1974) Operations research in production planning, scheduling and inventory control. Wiley, New York

Kacar NB, Irdem DF, Uzsoy R (2012) An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms. IEEE Trans Semicond Manuf 25(1):104–117

Kacar NB, Moench L, Uzsoy R (2013) Planning wafer starts using nonlinear clearing functions: a large-scale experiment. IEEE Trans Semicond Manuf 26(4):602–612

Kacar NB, Moench L, Uzsoy R (2016) Modelling cycle times in production planning models for wafer fabrication. IEEE Trans Semicond Manuf 29(2):153–167

Kefeli A (2011) Production planning models with clearing functions: dual behavior and applications. Unpublished Ph.D. Dissertation. Edward P. Fitts Department of Industrial and Systems Engineering. North Carolina State University, Raleigh, NC

Koltai T, Terlaky T (2000) The difference between the managerial and mathematical interpretation of sensitivity results in linear programming. Int J Prod Econ 65:257–274

Leachman RC (2001) Semiconductor production planning. In: Pardalos PM, Resende MGC (eds) Handbook of applied optimization. Oxford University Press, New York, pp 746–762

Leachman RC, Carmon TF (1992) On capacity modeling for production planning with alternative machine types. IIE Trans 24(4):62–72

Manne AS (1957) A note on the Modigliani-Hohn production smoothing model. Manag Sci 3(4):371–379

Missbauer H (2014) From cost-oriented input-output control to stochastic programming? Some reflections on the future development of order release planning models. In: Gössinger R, Zäpfel G (eds) Management Integrativer Leistungserstellung. Festschrift für Hans Corsten. Duncker & Humblot GmbH, Berlin, pp 525–544

Missbauer H, Uzsoy R (2011) Optimization models of production planning problems. In: Planning production and inventories in the extended enterprise: a state of the art handbook. Springer, Boston, pp 437–508

Modigliani F, Hohn FE (1955) Production planning over time and the nature of the expectation and planning horizon. Econometrica 23(1):46–66

Orlicky J (1975) Material requirements planning: the new way of life in production and inventory management. McGraw-Hill, New York

Pochet Y, Wolsey LA (2006) Production planning by mixed integer programming. Springer Science and Business Media, New York

Pürgstaller P, Missbauer H (2012) Rule-based vs. optimization-based order release in workload control: a simulation study of an MTO manufacturer. Int J Prod Econ 140:670–680

Rubin DS, Wagner HM (1990) Shadow prices: tips and traps for managers and instructors. Interfaces 20(4):150–157

Schneeweiss C (2003) Distributed decision making. Springer-Verlag, Berlin

Spitter JM, Hurkens CAJ, de Kok AG, Lenstra JK, Negenman EG (2005) Linear programming models with planned lead times for supply chain operations planning. Eur J Oper Res 163(3):706–720

Vollmann TE, Berry WL, Whybark DC, Jacobs FR (2005) Manufacturing planning and control for supply chain management. McGraw-Hill, New York

Voss S, Woodruff DL (2003) Introduction to computational optimization models for production planning in a supply chain. Springer, Berlin, New York

Voss S, Woodruff DL (2006) Introduction to computational optimization models for production planning in a supply chain. Springer, New York

Wight O (1970) Input-output control: a real handle on lead times. Prod Invent Manag J 11(3):9–31

Zipkin PH (2000) Foundations of inventory management. Burr Ridge, IL, Irwin