Chapter 16
Optimization Models of Production Planning Problems

Hubert Missbauer and Reha Uzsoy

16.1 Introduction

Mathematical programming formulations have been proposed for a wide range of production-related problems since the 1950s, addressing problems of long-term aggregate production planning, medium-term allocation of capacity to different products, lot sizing and product cycling, and detailed short-term production scheduling. The range of problems addressed by these methods spans a variety of managerial levels, problem environments, and time scales, often confusing even experienced practitioners and researchers as to what exactly is meant by the term "production planning." Hence, it is necessary to begin a chapter of this nature by specifying exactly what type of problems we plan to address, especially since an overview of all related areas is clearly beyond the scope of any single chapter. Indeed, we note in passing that there appears to have been no effort to collect all the production-related operations research literature in a single volume since the book by Johnson and Montgomery (1974), which remains an excellent reference for the basic concepts in the field.

In this chapter, we shall focus on manufacturing planning and control (MPC) systems for discrete parts manufacturing, where products are assembled from a variety of components, each of which in turn is produced by a multistage process. Manufacturing structures of this type occur in many industries, such as mechanical products, electrical appliances, electronics, and automotive manufacturing. Most facilities are not dedicated to specific products, and thus may require time-consuming changeovers when switching from one product type to another. Many insights in this chapter are also relevant for industries that only partly show these characteristics (e.g., semiconductor manufacturing). Process industries (e.g., steelmaking, continuous chemical processing or paper industry) often have substantially different

R. Uzsoy (✉)
Edward P. Fitts Department of Industrial and Systems Engineering,
North Carolina State University, Raleigh, NC 27695-7906, USA
e-mail: ruzsoy@ncsu.edu

K.G. Kempf et al. (eds.), *Planning Production and Inventories in the Extended Enterprise*,
International Series in Operations Research & Management Science 151,
DOI 10.1007/978-1-4419-6485-4_16,

characteristics, so we will not address these cases. For a discussion of MPC systems in process industries, see Gunther and Van Beek (2003); for steel plants as a special case, see Tang et al. (2001) and Missbauer et al. (forthcoming).

For the types of manufacturing systems under consideration, a planning logic has evolved over the last 45 years starting from bill-of-material explosion and leading to the Material Requirements Planning (MRP) (Orlicky 1975) and Manufacturing Resource Planning (MRPII) systems (Wight 1983), which form the basis for most of the MPC and supply chain planning systems in industrial use today (Vollmann et al. 2005). Today's Advanced Planning and Scheduling (APS) Systems (Stadtler and Kilger 2008) aim at complementing MPC systems by concentrating on planning and coordinating the material flow between companies or manufacturing plants, leveraging the data collection and organization capabilities of the Enterprise Resource Planning (ERP) and Manufacturing Execution Systems (MES) used by many companies today. The chapter by Fordyce et al. in this volume gives an excellent view of such a system.

The basic problem of production planning in these environments is essentially one of matching supply to demand. This involves viewing the production system as a network of resource groups, which we shall refer to as work centers, and allocating the capacity of production resources at these work centers among different products over time, coordinating the associated inventories and raw material inputs so that known or predicted customer demand is met in the best possible manner. The "best possible manner," of course, requires more precise definition in order to form the basis of an optimization model, and can vary widely based on the specific production environment being considered. However, the objective functions are generally aimed at minimizing the total expected costs of production and inventories over the time horizon considered. There are also a number of models where demand is influenced by management decisions such as pricing, in which case the objective takes the form of profit maximization.

The decision variables emerging from the production planning activity depend on the decision structure of the MPC system, mainly on the extent to which detailed (mainly scheduling) decisions are made at the planning level. In the following, we assume that detailed scheduling decisions are made on the shop-floor level. This is the usual MPC structure and is described below in more detail. Given this structure of the MPC system, the basic decision variables emerging from the production planning activity are the amount of material (work orders) for each type of product that is to be released to each production resource over time, together with the required due dates. We shall focus on these decisions throughout this chapter since ultimately these decisions are the only actionable ones resulting from the production planning process. Estimates of a number of other quantities, such as production quantities at each production work center over time and inventory levels over time, are also obtained from production planning models, but we will take the position that all these ancillary quantities result from applying the work release decisions to the constraints defining the operation of the production resources (capacity constraints) and the material flows through the production system. We will assume that

capacity determination over the long term, the domain of the classical literature on capacity expansion (represented by, e.g., Luss 1982) is not part of this problem domain.

Due to the complexity of the supply chain, the manufacturing process and the organization responsible for managing and coordinating them, the MPC task is usually structured hierarchically (Anthony 1966; Bitran et al. 1981, 1982; Bitran and Tirupati 1993; Hax and Candea 1984). Essentially two planning levels can be distinguished as follows:

- *Upper level (central MPC):* The complexity of the supply chain or manufacturing process is generally addressed by aggregating segments of the production process into departments or production units (Bertrand et al. 1990). This level then involves planning of the material flow over the entire logistic chain at an appropriate level of aggregation, without determining detailed schedules (production sequences) within the production units.
- *Lower level:* Detailed scheduling of the work orders within the production units, which involves determining the start and finish dates of the operations and their production sequence at the facilities.

Our specification of these levels follows the use of the terms "goods flow control" and "production unit control" in (Bertrand et al. 1990). Note that the upper level frequently is hierarchical in itself; see Vollmann, Berry et al. (2005), for its structure in today's MPC systems, and Bertrand et al. (1990) and de Kok and Fransoo (2003), for advanced planning architectures. Many of these concepts date back at least to the book by Anthony (1966), and an extensive literature has been developed around these issues (for conceptual issues of hierarchical production planning, see Schneeweiß 2003). When a hierarchical MPC system is designed, the data and the decision variables are aggregated with respect to products (aggregating products into product groups or families), capacity (aggregating resources to resource groups), and time (Stadtler 1996). Decisions taking place at a similar level of detail – hence often over similar time intervals or frequencies – are considered together in the same level of the hierarchy, and a mechanism is devised to propagate the implications of these decisions up and down the hierarchy. However, the two levels outlined above will be sufficient to motivate the work we wish to accomplish in this chapter.

The essential interface between the two planning levels is *order release*. The central MPC system coordinating the production units issues *work orders* that specify the particular product or component type to be produced, the amount to produce, and the due dates of the orders. Control over the work order then passes to the lower, detailed scheduling level within the production unit concerned. Clearly this approach requires *coordination norms* between the planning levels, which is most commonly implemented using *planned lead times*. The upper level creates its plans based on some assumptions as to the lead time, the time between the order being released into production and its being completed. The production unit must behave in a manner consistent with this assumption. The most common approach is for the upper level to assume a constant lead time for the released orders, which the

production unit commits to meeting, at least as long as capacity utilization remains within reasonable limits. Thus the management of lead times by the production unit, and of the norms defining the upper level's view of the defined lead times, are of utmost importance.

In any hierarchical planning system, the upper level can evaluate its decision alternatives accurately only if it has access to a model that predicts the behavior of the system controlled by the lower level. Thus this *anticipation function* of the upper level (Schneeweiß 2003, p. 33 ff.) is a crucial element, allowing the upper level to anticipate the consequences of its decisions for the lower level(s). In our case, the central MPC system at the upper level needs a model that predicts the performance indicators of the manufacturing system (work in process inventory, flow time, due-date performance, etc.) resulting from an order release plan. *The formulation of mathematical programming models for the upper level of the MPC system described above that use a variety of different anticipation functions is the central topic of this chapter.*

Perhaps the most obvious consequence of shortcomings in the design of the upper level of the MPC system with respect to the anticipation of the dynamic behavior of the manufacturing system is insufficient management of lead times and work-in-process (WIP). In both Europe and the USA, a considerable body of work has approached these issues using the terminology of workload control. This type of approach usually implements the hierarchy discussed above, with the upper level (central MPC) performing order acceptance and deciding what demand will be served and which demand will be delayed in order to balance load and capacity in the medium term. The tradeoff between low flow times and high throughput is managed by the release of work into the production units over the short term based on some statistics reflecting the state of the production system.

In the following, we describe the workload control concept that provides the basis for Order Review and Release (ORR) systems that usually do not incorporate optimization models. In the subsequent sections we concentrate on optimization models for planning the aggregate material flow and order release in MPC systems.

16.2 Nonoptimization Workload Control

Since most of the MPC systems in industrial use today have largely evolved in practice (see McKay chapter in this volume) and have not been designed from an underlying theoretical basis, it is not surprising that their approaches to lead time management are often problematic. A well-known example is the *lead time syndrome* (Wight 1983, p. 108 ff.) that in extreme cases can inflate lead times beyond any arguable level. This can occur when lead time is considered to be a *forecast* variable, which is the usual practice in MRP systems: Lead times are estimated from realized values from the past, rather than being determined by the state of the production system. It is a crucial insight that lead times are *workload-dependent* and thus should be regarded as *control* variables (for this comparison, see Tatsiopoulos

and Kingsman 1983): Lead times are determined mainly by the waiting times at the work centers which are determined by the utilization level of the resources, which are determined by the amount of released work. This insight, supported by a very extensive literature on queuing models of manufacturing systems (Hopp and Spearman 2001; Buzacott and Shanthikumar 1993), motivates the workload control (WLC) concept as a way to improve lead time management.

The workload control (WLC) concept considers flow times as output variables that can be controlled by the manner in which work is released into the shop over time. If the amount of work released (measured, e.g., in standard hours) is large, this leads to longer queues at the work centers and hence to longer waiting times and flow times. Reliable flow times can only be maintained if the amount of work released into the system and its output (determined by the available capacity) are balanced such that the queues of work in process inventory (WIP) at the work centers, usually measured in hours of work, are kept at a predefined level. This makes *order release* an essential decision function and a core component of WLC. Implementing any form of WLC thus requires the solution of two subproblems: (1) determining the target WIP level and (2) determining the release dates of orders, i.e., how work will be released into the shop over time.

The target WIP level must be a compromise between the goals of maintaining low WIP level and short flow times on the one hand and high output on the other. A high output requires an average level of WIP (queues at the work centers) that prevents idleness of the work centers, buffering against *variability* in the material flow between stations (see Hopp and Spearman 2001, p. 287 ff., for the "corrupting influence of variability"). This can be formalized as a functional relationship: Once the average WIP level is determined, the average flow time and the output (capacity utilization) are also determined, given the order and shop characteristics. The functional relationships between average WIP and other important performance indicators are usually expressed as *characteristic curves* that are often determined by simulation. Figure 16.1 shows an example. For this figure the manufacturing system has been simulated several times with different average levels of WIP in the shop. The average WIP levels were obtained by simulating the load-oriented order release approach of Wiendahl (1995) with different target WIP levels. The realized mean WIP values are shown on the x-axis, and the mean flow time and output on the y-axis.

It is important to note that the characteristic curves in Figure 16.1 are not completely determined by the technical properties of products and manufacturing system. They also depend on sequencing rules, lot sizes, capacity flexibility, order release frequency, etc., that is, they are a result of the long-term characteristics of the planning system. Thus WLC can be considered as a means of *complexity reduction*: the specification of long-term decision rules in the MPC system leads to stable operational characteristics of the manufacturing system it controls (expressed as characteristic curves), which, in turn, provide the basis for order release decisions that maintain the desired flow times. Consequently, WLC is an architecture for the entire MPC system (for conceptual issues, see Bertrand et al. 1990; Zäpfel and Missbauer 1993b; Missbauer 1998).

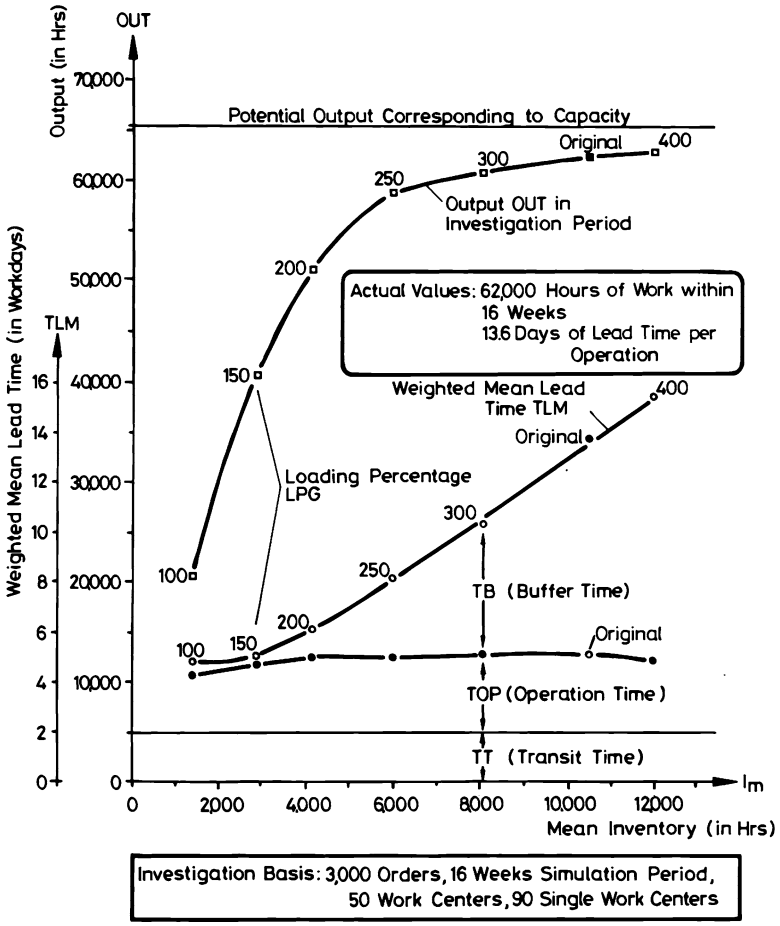

Fig. 16.1 Curves for output and mean flow time per operation with change in the work-in-process Wiendahl (1995, p. 246)

The decision problem of determining how to release work over time to maintain a predetermined target WIP level has been a research topic since the 1970s. Two main approaches can be distinguished as follows:

(a) WLC order release mechanisms that determine the work orders to release for a short planning horizon. We use the term "traditional WLC order release mechanisms" since this has been an essential part of WLC research in the last 20 years (Land 2004; Stevenson and Hendry 2006). We describe this general approach and its limitations in the following subsection.

(b) Order release planning procedures based on an explicit model of the material flow and the time-dependent WIP level over a longer planning horizon, usually divided into periods. This approach, its present state, and research topics are

described in Sects. 16.3 and 16.4. The focus of this chapter is very much on order release methods that are explicitly based on the WLC concept and that are applicable for complex product/process characteristics.

16.2.1 Traditional Order Release Methods Based on Workload Control

Based on the insights described above, a number of short-term order release mechanisms have been designed, mainly in the 1980s and 1990s. Examples are CONWIP (Spearman et al. 1990), load-oriented order release (Wiendahl 1995), and the order release stage of the LUMS approach (Hendry and Kingsman 1991; Stevenson and Hendry 2006) and the method of Bertrand and Wortmann (1981). In the following, we describe the general concept of these methods. Overviews can be found in Bergamaschi et al. (1997), Land (2004) and Fredendall et al. (2010). Kanban, which can also be considered as a Workload Control technique, is based on more restrictive assumptions and is not included here. Drum-Buffer-Rope is based on a more detailed schedule of the bottlenecks ("drum beat") and thus is different from the hierarchical MPC concept that provides the basis for this chapter (for a discussion of MPC concepts, see Zäpfel and Missbauer 1993a; for Drum-Buffer-Rope, see Cohen 1988; Gupta 2005).

The order release mechanisms described in this section usually are based on the following situation: The work orders in the MPC-system are generated from customer orders or from the MRP system from net requirements, lot sizing, and infinite capacity loading. Since in WLC order release is a decision function, the work orders are initially held in an *order pool* of unreleased orders. These orders are specified by product or component type, lot size, and required due date. A planned start date, derived from the required due date and a planned flow time that is consistent with the WIP norm, is usually available. Orders from the pool are released according to their planned start dates and the load situation in the shop. All orders should be finished on time, and the WIP norms should be maintained. When an order is released, control over the order is transferred to the scheduling level of the production unit, which has to meet the required due date.

We define a *traditional WLC order release mechanism* as follows (see also the basic release procedure described in Land (2004, p. 36 ff.)); we roughly follow the classification framework in (Bergamaschi et al. 1997). Order release is load limited, that is, a load limit is determined for each work center or for the shop as a whole, based on the characteristic curves of Fig. 16.1, and the order release mechanism prevents the load limits from being exceeded (in some cases with minor tolerance). The planning horizon is one period (in practice, usually 1 day to 1 week), and there is usually no order release plan for multiple periods. The capacities of the work centers are usually fixed, although some mechanisms consider capacity adjustments. The timing of order release may be event-driven in continuous time, or periodic at

the beginning of each planning period; both options can be used simultaneously. An order is released if its release does not violate these workload norms, defined as upper and/or lower bounds. The subset of orders to be released within these constraints is usually selected heuristically (e.g., according to a priority following the urgency of the order), but integer programming can also be used, as in, for example, Irastorza and Deane (1974). When order release is performed, the feasibility of releasing the selected orders is checked against the load limits. This feasibility check requires the definition of how WIP is to be measured (e.g., in number of orders or in hours of work) and what WIP is to be considered in the decision (WIP for the entire shop, for each work center separately or just for the bottleneck work centers). Many of these issues of how to aggregate and measure WIP will also arise in our discussion of clearing functions in Sect. 16.5.

If a target WIP level is defined for individual work centers, then controlling the load at the work centers (*direct load*) is the most detailed technique. The future value of this direct load has to be estimated because the work input to the WIP level at a work center is controlled by sequencing decisions at the upstream work centers and is not known at the time of order release. Alternatively, only the *aggregate load* of the work centers, defined as the released work in the shop that has to be processed by a work center irrespective of its current position, is controlled, which avoids estimation of the order arrival patterns at the work centers.

When an order release mechanism is designed, the design options mentioned above must be specified. Figure 16.2 summarizes a well-known classification of these design options.

Once the order release mechanism has been designed by specifying the design options discussed above, its realized performance is determined by the parameter settings. The most important parameters are as follows:

- Target WIP (WIP norms) at the work centers. This parameter, which is closely related to the target value of the average flow time, can be expressed in different ways depending on the order release mechanism (e.g., load limit or load percentage in load-oriented order release) and specifies the compromise between output and WIP-flow time.
- A time limit that prevents the premature release of orders whose planned start date is far in the future. This parameter plays an important role in determining the extent of production smoothing (see below).

Other parameters are the length of the planning period and the order release frequency if order release is performed periodically (Perona and Portioli 1998).

Most of the order release mechanisms developed since 1980 follow this logic and can be obtained by specifying the design options above. We call these methods traditional because they share a common structure and have formed the mainstream of the research on WLC-based order release mechanisms in the last 25 years. They have also been tested extensively and are partly available in standard software. A stream of such methods and their interactions with shop-floor dispatching has been explored by a body of work focusing on the semiconductor industry,

Dimensions	Options
Order release mechanism	Load limited Time phased
Timing convention	Continuous Discrete
Workload measure	Number of Jobs Work quantity
Aggregation of workload measure	Total shop load Bottleneck load Load by each workcentre
Workload accounting over time	Atemporal Probabilistic Time bucketing
Workload Control	Upper bound only Lower bound only Upper and lower bounds Workload balancing
Capacity planning	Active Passive
Schedule visibility	Limited Extended

Fig. 16.2 Design options in traditional order release mechanisms (Bergamaschi et al. 1997)

as discussed in Uzsoy et al. (1994). For a comprehensive description of order review/release methods, which is beyond the scope of this chapter, see Land (2004), Bergamaschi et al. (1997) and Philipoom and Fry (1992).

Order release mechanisms of this type have a short planning horizon, but the release decisions determine the order arrivals at the work centers for a period that equals the flow times of the orders and also influence the options available to future release. Thus it is important to gain insight in how order release mechanisms work and how the goal of maintaining a stable WIP level can be achieved over a longer time horizon. Three topics are relevant here as follows:

- Order release determines the start dates of the orders and thus the extent of workload smoothing if (e.g., seasonal) demand variations occur.
- The workload norms must be consistent with the desired output, which in turn can depend on the extent of workload smoothing.
- Especially in the case of multiple products with different routings and resource requirements, the order release mechanisms should be able to perform load balancing among work centers. That is, the sequence of order releases should avoid temporary over- or underloading of certain routes or work centers. This is

essential for achieving high throughput and low WIP at the same time without just shifting the waiting time to the job pool.[1]

These tasks require the capability to look ahead beyond the short planning horizon of the release mechanisms. There are two ways to accomplish this as follows:

- Short-term release mechanisms can be complemented by a medium-term planning level that balances load and capacity and thus determines the required output over time. The release mechanisms perform short-term control that keeps WIP and shop flow times under control. This shifts much of the planning task to the medium-term level, and especially if flow times are long (e.g., semiconductor manufacturing with hundreds of operations per order), integration of medium-term planning and short-term order release can be difficult.
- The dynamic behavior of order release mechanisms and the material flow that results from release decisions can be controlled by the design of the release mechanisms and by the parameter setting (especially WIP norms and time limit). For instance, if load leveling over a longer horizon is desired, the time limit should be long, which allows early release of orders in periods with low demand. If utilization is low, the time limit should be short, because this prevents early release and completion of orders, etc. (see Zäpfel et al. 1992). Research, mainly by simulation, has accumulated a body of knowledge on decision rules that provide support for design and parameter setting of release mechanisms for specified material flow structure, demand pattern, etc. (for reviews, see Land 2004; Stevenson and Hendry 2006). Hence the traditional order release mechanisms can be regarded as a rule-based approach to optimize the material flow through a production unit as discussed in Missbauer (2009).

In the following sections we focus on optimization models that determine the optimal aggregate material flow through a production unit, assuming that demand forecasts are available at least at an aggregate level (groups of products). This optimized aggregate material flow forms the basis for order release that leads to this planned material flow. It can be expected that the potential of optimization models that determine order release is higher than the potential of release mechanisms that are controlled by appropriate parameter setting. If no reliable demand forecasts are available (especially in the case of customer order-driven production), order release planning might be difficult. In this case, the essential problem is to coordinate customer enquiry/order acceptance and order release/WIP control (for this topic, see Hendry and Kingsman 1991; Stevenson and Hendry 2006). The planning models described in the remainder of the chapter perform many of the functions of

[1] Simulation results indicate that traditional order release mechanisms can in fact reduce average *total* throughput time of orders (pool waiting time plus shop flow time) and do not just shift the waiting time from the shop to the pool, possibly increasing total flow time due to reduced shop capacity (for this critique, see Kanet 1988). A possible reason is the load balancing effect. A good order release mechanism limits the shop load, but it also aims at keeping the level of WIP at the target level for each work center, thus balancing the load among the work centers by releasing orders with different routing. (For simulation results on total throughput time reduction, see Land 2004.)

order release mechanisms (WIP control, load leveling, load balancing) and thus can potentially replace the release mechanisms to a large extent. However, depending on the level of detail of the models, implementation of the order release plans by means of a traditional release mechanism may still be necessary or useful. In this case, the traditional order review and release mechanisms discussed in this section will operate in the very short term as part of the lower, detailed scheduling level of the MPC hierarchy described in Sect. 16.1. When applied in this way, the release mechanisms take as input the quantity of work that is required to be released over some planning period, say a week or a month, and control releases to the shop floor based on the state of the system in close to real time. Applying multiperiod optimization models to determine optimal time-varying parameter settings for release mechanisms (Zäpfel and Missbauer 1993a) might be a serious alternative but remains largely a topic for future research.

In the following section, we give a brief historical review of the development of optimization models for aggregate material flow planning in MPC systems.

16.3 Historical Review

While production planning has obviously been executed in some form since the beginning of even craft production, the development of quantitative methods for these problems has surprisingly of recent origins. The chapter by McKay in this volume gives an overview of the historical development of production planning since the beginning of the Industrial Revolution. While the well-known work of Harris (1915) launched the area of inventory modeling, it was not until the 1950s that this became a major research area. Similarly, the use of optimization models for planning production over time has its origins in the work of Modigliani and Hohn (1955), although the roots of this work reach back to the activity-based economic models developed by economists such as Leontief and Koopmans (e.g., Koopmans et al. 1951). The area of sequencing and scheduling also had its pioneering papers in this period, notably those of Jackson (1955) on single machine scheduling problems and Manne (1960) on job shop scheduling. However, we will focus the discussion here on production planning, leaving detailed discussion of the extensive field of sequencing and scheduling to specialized volumes in that area (Pinedo 1995; Pinedo and Chao 2005; Parker 1995).

It is interesting to note that the basic formulations of most classical production planning-related problems were essentially in place by 1960, when the book by Holt et al. (1960) collecting their previous work appeared. Hence, it is worthwhile examining these early papers in some detail to gain insight into why these formulations developed the way they did, and what the implications of these are relative to the current situation.

The work of Modigliani and Hohn (1955) views the problem of production planning over time as that of trading off production costs against inventory holding costs. Production costs are assumed to be convex, with increasing marginal production

costs, while inventory holding costs are approximated by the time average of the ending period inventories, leading to a linear cost function. The problem is formulated on discrete time periods, the cost function is assumed to be stationary over time, demand in each period is known with certainty, and no backlogging is allowed. The monotone increasing marginal production cost makes it more economical to meet periods of high demand by producing in prior periods of low demand and holding inventory, giving the basic tradeoff in the problem. The authors develop an optimal solution based on calculus that essentially identifies planning horizons, allowing the problem to be decomposed along the time horizon into subproblems consisting of a certain number of consecutive periods that can be solved independently. This approach forms the basis for the later work of Holt et al. (1955, 1956), which subsequently led to the HMMS book (Holt et al. 1960). It is also interesting that in Chap. 6 of their book they explicitly address the extension of their decision rules to an environment with uncertain demand, and show that under a quadratic objective function of the type they assume, the deterministic equivalent of the stochastic problem is achieved by using expected demand values in their deterministic rule, which is equivalent to assuming an unbiased demand forecasting procedure. This insight appears to have motivated the heavy focus on deterministic models, although the proof they provide is valid for the specific case of a quadratic objective function. An interesting discussion of this body of work is given by Singhal and Singhal (2007).

A number of interesting points emerge from this work. It is interesting that capacity constraints are not modeled; the implicit assumption appears to be that capacity can be varied in the short run, and the costs of doing this contribute to the increasing marginal cost of production. This discussion is made more explicit in the context of labor costs by Charnes et al. (1955). It is also interesting that while the cost function is explicitly built up from holding production and fixed costs independent of production volume there is no discussion of how one might actually estimate these costs from existing business records. Finally, the basic paradigm is that of modeling the physical flows in the problem – production and inventories – and assigning costs to these, rather than modeling the cash flows explicitly as a means of capturing the financial impact. This paper also seems to have motivated the idea that in problems over time, perfect information of the entire planning horizon is not necessary, but rather just the first few periods on a rolling horizon basis is quite close to optimality. This has led to a long stream of papers using these and related ideas, including the well-known dynamic lot sizing model of Wagner and Whitin (1958).

In the mid-1950s, it began to be realized that the emerging linear programming technology could be applied to production planning problems quite directly. In particular, researchers realized that models of the type addressed by Modigliani and Hohn (1955) and Holt et al. (1956) could be formulated as linear programs. The two principal papers that appear to have accomplished this independently of each other are Hanssmann and Hess (1960), whose title very much resonates with the HMMS work, and Manne (1957). Another notable early paper is that of Bowman (1956), which appears to be one of the earliest to identify the extensive presence of network structure in production planning models.

At this point, the principal characteristics of the mathematical programming models used for production planning problems were now in place. The decisions cover a planning horizon that is divided into a number of discrete time periods, each of which has an associated set of decision variables reflecting the decisions made in that period. The decision variables represent the physical flows of material through the different production resources; the objective function is generally that of minimizing the variable costs of production, inventories and backlogs over the planning horizon, and capacity constraints on the production resources in each period are satisfied at an aggregate level. In the following section we shall investigate the basic assumptions of these models in more detail, focusing on the manner in which they model the dynamics of capacitated production resources.

16.4 Optimization, Planning, and Work Release

It is interesting to note that although the decisions made in well-known optimization models are referred to almost uniformly as "production" decisions, in reality these models usually are work release models; in hierarchical manufacturing planning and control systems as described in Sect. 16.1, the only way they can be implemented is through the release of work orders with specified due dates into the production facility being modeled. As such, these models are closely related to the extensive stream of work on work release, order review/release (ORR), and workload control discussed in the previous section. They also address the basic problem of planning time-phased work releases addressed by the well-known and extensively implemented Material Requirements Planning (MRP) procedure (Vollmann et al. 1988; Baker 1993) and related techniques. Finally, it is worth noting that these problems have also been addressed in several research streams out of the artificial intelligence community (e.g., Smith 1993; Zweben and Fox 1994).

The vast majority of the mathematical programming models of interest to this chapter approach the problem in the same manner. The time horizon being considered 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 to minimize the total cost, which may be defined in different ways depending on the specific environment being considered, over the planning horizon. Following Hackman and Leachman (1989), we can view these models as containing three basic sets of constraints:

1. *Inventory or material balance equations*, which capture the flows of material through both space and time. These will also enforce the satisfaction of demand, which is viewed as a flow of material from 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 first two sets of constraints are critical to the accurate reflection of the actual behavior of the production system, which in turn is essential to the optimality and feasibility of the production plans obtained from the model. Following the discussion in Sect. 16.1, their aggregate nature requires these models to use some type of anticipation function that predicts the effect of their decisions on the detailed scheduling level of shop operation. As discussed in Sect. 16.1, the manner in which production lead times (also referred to as *cycle times* or *flow times*)[2] are treated is a crucial aspect of this anticipation function, affecting both the capacity and flow balance constraints above. We now examine several different models of lead times used in production planning models, starting from the simplest, and discuss their advantages and disadvantages.

16.4.1 Fixed Delays Independent of Workload

The simplest model of lead times that is encountered quite commonly in inventory models is one of instantaneous replenishment where the quantity ordered at a given point in time becomes available immediately upon ordering. The closest equivalent to this in the domain of mathematical programming models is to assume that lead times are approximately equal to one period, i.e., that the quantity of material R_t that is released into the system at the start of period t is available to meet demand at the end of that period. In order to ensure continuity of the solution between periods, we need to model the flows of material in and out of the finished goods inventory, which yields the system dynamics equations

$$I_t = I_{t-1} + R_t - D_t \tag{16.1}$$

where I_t denotes the finished goods available on hand at the end of period t and D_t the demand at the end of that period.

However, this is clearly often not realistic, and a more commonly encountered model in both inventory and the mathematical programming literature is a fixed, deterministic replenishment lead time that is independent of the quantity ordered or released. The stochastic equivalent of this model is a random lead time with a time-stationary probability distribution that is independent of the order quantities, such as the case treated by Eppen and Martin (1988), or the class of models discussed in Chap. 7 of Zipkin (1997). In this case, the amount Q_t ordered at the start of period t becomes available at the beginning of period $t + \tau$, where the integer τ denotes the fixed lead time. In a production system, the amount R_t released into the system at the start of period t becomes available for use at the start of period $t + \tau$. If we denote

[2] In general, we use the term *flow time* when we consider this time span from a manufacturing perspective, and the term *lead time* when it is considered from a planning perspective; see Hopp and Spearman (2001, p. 321). The terms are not always clearly distinguished in the literature.

the output of the production system in period t by X_t, we have the relationship $X_t = R_{t-\tau}$. The system dynamics are now described by the relationship

$$I_t = I_{t-1} + X_t - D_t = I_{t-1} + R_{t-\tau} - D_t. \tag{16.2}$$

We note in passing that this is exactly the model of lead times used in MRP in its backward scheduling phase. It is common both in the literature and in practice to assume that the fixed lead time τ corresponds to an integer number of planning periods; Hackman and Leachman (1989) present a straightforward method for managing lead times that correspond to a fractional number of planning periods, which we will discuss later.

The difficulty with this model in the context of production systems is that it assumes that there is no limit on the amount of material the system can produce in the given lead time. Hence, most optimization models of capacitated production systems will limit the total output of the system in a given period by imposing a constraint of the form $X_t \leq C_t$, where C_t denotes the maximum possible output of the production system in a given period. For exposition, let us assume that each unit produced requires one time unit of the resource, and the resource capacity is expressed in terms of time units available per planning period. There is now the question of reconciling the capacity constraint with the system dynamics constraint, which involves determining at what point in time releases into the system in period t occupy the capacity of the resource. Three logical constructions can be distinguished here: (1) "lag before" models that assume that the resource capacity is occupied at the end of the lead time of this operation, (2) "lag after" models where the resource capacity is occupied at the beginning of the lead time, and (3) models that allocate the production date (and the resource requirements) within the lead time and hence allow production smoothing within the lead time. For "lag before" and "lag after" models we refer to Hackman and Leachman (1989); models of type (3) are formulated in de Kok and Fransoo (2003). For the discussion of the treatment of WIP, in the following we describe the "lag before" models that assume that the releases R_t in period t occupy the resource capacity in the period that the output is produced, implying $X_t = min\{R_{t-\tau}, C_t\}$. This corresponds to the "lag before" models in Hackman and Leachman (1989).

We can thus describe the behavior of this system in a given period t with the set of constraints

$$\begin{aligned} X_t &= R_{t-\tau}, \\ I_t &= I_{t-1} + X_t - D_t, \\ X_t &\leq C_t. \end{aligned} \tag{16.3}$$

The first constraint is explicitly included for exposition; clearly in practice we would make the substitution to eliminate one of the two variables from the formulation.

Most common LP models for production planning will refer to "production" variables X_t, but these generally correspond to releases into the production system, with resource capacity being occupied τ periods after the release has taken place. Note that the amount of production that can take place in a given period is limited by both the capacity C_t and the amount of work available for processing, given by past releases $R_{t-\tau}$ per the first constraint. Hence the amount of WIP W_t available for the resource to process in period t is simply $R_{t-\tau}$. However, if we define WIP as the inventory literature defines on-order inventory, as orders that have been released but not yet completed, at the end of period t the production system in this model will have a WIP level of

$$W_t = \sum_{k=t-\tau+1}^{t} R_k - \sum_{k=t+1}^{t+\tau} X_k \tag{16.4}$$

units of product. It is interesting to note that this quantity does not generally appear in the constraints or objective functions of most common LP models of production planning, although as seen above it is not difficult to model; an exception is the model of Riaño et al. (2003). It is also interesting to note that only a portion of this WIP, given by $R_{t-\tau}$, is actually available to the resource for processing in period t.

The deficiency of this model is rooted in its treatment of WIP. Essentially it assumes that WIP will not accumulate in the system over time; the releases in period $t - \tau$ constitute the entire WIP available to the resource for processing in period t. The releases are implicitly constrained not to exceed the capacity, so the system is always able to process all its available WIP in a single period. The remainder of the WIP, given by

$$\sum_{k=t-\tau+2}^{t} R_k, \tag{16.5}$$

has no effect on the cycle time of the resource, which is always equal to the prespecified parameter τ, and, as far as this model of production capacity goes, is completely unrelated to the capacity C_t of the resource in a given period. All the lead time τ accomplishes is to delay the arrival of work at the resource after its release into the system; it does not describe the behavior of the resource itself, which is assumed in the capacity constraint to be able to process any amount of material up to the capacity limit C_t in a given period. This also explains an interesting anomaly with this type of model that positive dual prices for capacity constraints result only when the capacity is fully utilized. However, queuing models repeatedly show that system performance, especially as related to WIP levels and cycle times, often begins to degrade at utilizations substantially below 1, implying the existence of situations where even though a resource is not fully utilized, additional capacity at the resource might be beneficial to system performance, although adding that capacity would not necessarily be economically desirable.

To summarize this discussion, the conventional view of production capacity used in MRP and most mathematical programming models results in an LP of the following form:

$$\text{minimize} \sum_{t=1}^{t} (h_t I_t + c_t R_t) \tag{16.6}$$

subject to

$$I_t = I_{t-1} + R_{t-\tau} - D_t, \quad \text{for } t = 1, \ldots, T, \tag{16.7}$$

$$R_{t-\tau} \leq C_t, \quad \text{for } t = 1, \ldots, T, \tag{16.8}$$

$$R_t, I_t \geq 0, \quad \text{for all } t = 1, \ldots, T. \tag{16.9}$$

As pointed out by Hackman and Leachman (1989), most LP models encountered in practice will involve additional constraints specific to the application domain under study, but the model above represents the essentials of inventory balance between periods and aggregate capacity within periods. Note that because all rates are uniformly distributed over a planning period, ensuring nonnegative inventory levels at the boundaries between periods is sufficient to ensure inventory is nonnegative throughout the period.

Models that involve lot-sizing considerations may involve integer variables, but the basic view of system capacity and lead times is usually not very different from this. We have chosen a simple objective function, that of minimizing the sum of production and inventory holding costs over the planning horizon. Clearly far more elaborate objective functions are possible, but our emphasis in this chapter is on the representation of production capacity and system dynamics. Finally, we note that backlogging of any form is not allowed, again in service of our focus on the ability of different sets of constraints to accurately represent the actual capabilities of the production system.

We now examine some other approaches that have extended this basic model without fundamentally altering its treatment of lead times.

16.4.2 Formulations Based on Lead Time Distributions

The basic assumption of a fixed lead time equal to an integer number of planning periods has been relaxed by a number of authors. Leachman and his coworkers (Hackman and Leachman 1989; Hung and Leachman 1996) have proposed models where the lead time associated with each planning period can be a fractional number of planning periods. Their approach is essentially equivalent to assuming a deterministic lead time distribution for the input in period t, which specifies the fraction $w_{it\tau}$ of the amount of product i released in period t that will emerge as finished product in period τ i.e., will have a lead time of $\tau - t$. An alternative approach

is to consider the $w_{it\tau}$ values as random variables whose probability distribution depends on the state of the production system. We shall begin our discussion with the workload-independent approach of Hackman and Leachman (1989), which has formed the basis for several industrial implementations in the semiconductor industry (Leachman 1993; Leachman et al. 1996) and a number of other refinements (Kim and Leachman 1994; Kim et al. 1996; Dessouky and Leachman 1997). A number of authors have taken related approaches based on the use of planned lead times (Spitter et al. 2005a, b). We shall then discuss the workload-dependent model of (Riaño 2003; Riaño et al. 2003), and some related approaches (Voss and Woodruff 2003; Lautenschläger and Stadtler 1998). The recent review paper by Pahl et al. (2005) is also a good source for some of this material, in addition to its discussion of the clearing function related methods presented in the next section.

16.4.3 Workload-Independent Lead Time Distributions

In order to present the basic approach to modeling workload-independent lead time distributions, we shall use the work of Hung and Leachman (1996) as the focus of the discussion, although the approach was originally proposed by Hackman and Leachman (1989). The basic formulation used is essentially the Step-Separated formulation of Leachman and Carmon (1992), which requires estimated lead times L_{ij} required for a lot of product i to reach operation j after being released into the plant. However, instead of fixed lead times that remain constant over the entire planning horizon, the authors associate values of the lead time parameters with the start of each planning period. In the following $t = 0$ is the start of period 1, $t = 1$ is the start of period 2, etc., that is, a time unit is the period length. The lead time parameters are defined as follows:

L_{ijt} Lead time required for a lot of product i to reach operation j if the lot reaches operation j at the end of period t (= at time t).

These lead times are allowed to take on fractional values. It is interesting to think how one might estimate these time-dependent lead time parameters in the absence of a production plan that defines the workload of the production resources over time. The iterative technique of Hung and Leachman (1996) discussed below addresses this issue directly.

Given these lead times, the loading of the production resource in period t is defined by releases occurring in the time interval $Q = [(t-1) - L_{i,j,t-1}, t - L_{ijt}]$, assuming planning period t starts at time $(t-1)$. There are two cases to consider here. In the first, simpler case, the time interval Q lies within a single planning period $\lceil (t-1) - L_{i,j,t-1} \rceil = \lceil t - L_{ijt} \rceil$ where the $\lceil x \rceil$ notation denotes the smallest integer greater than or equal to x. In this case, the proportional share of the amount released in period $\lceil (t-1) - L_{i,j,t-1} \rceil$ arrives at operation j in period t. This is consistent with the basic assumption of linear programming models that

activity intensities are uniformly distributed over the planning period, as discussed by (Hackman and Leachman 1989). Hence the amount Y_{ijt} of product i loading resources at operation j in period t is given by

$$Y_{ijt} = \left(\frac{(t - L_{ijt}) - \big((t-1) - L_{i,j,t-1}\big)}{\Delta} \right) e_{ijt} R_{i,\lceil (t-1) - L_{i,j,t-1} \rceil} \tag{16.10}$$

where Δ denotes the period length (which we set to 1 by definition), e_{ijt} denotes the overall yield of product i from the start of the process to step j in period t. If, on the other hand, the time interval Q spans multiple planning periods, we allocate the load due to releases in that period in proportion to the fraction of that period's total duration included in the interval Q (again assuming uniform release rates within the planning periods). This yields

$$\begin{aligned} Y_{ijt} = & \left(\frac{\big(\lceil (t-1) - L_{i,j,t-1} \big)\rceil - ((t-1) - L_{i,j,t} - 1)}{\Delta} \right) e_{ijt} R_{i,\lceil (t-1) - L_{i,j,t-1} \rceil} \\ & + \sum_{\tau = \lceil (t-1) - L_{i,j,t} - 1 \rceil + 1}^{\lceil t - L_{ijt-1} \rceil} e_{ij\tau} R_{i\tau} \\ & + \left(\frac{(t - L_{ijt})}{\Delta} \lceil t - L_{ijt-1} \rceil \right) e_{ijt} R_{i,\lceil t - L_{ijt} \rceil} \end{aligned} \tag{16.11}$$

The operation of this approach is illustrated in Fig. 16.3, from Hung and Leachman (1996). The upper part of the figure shows the uniform release rates in each planning period, while the lower graph shows the resource loading that results from these releases arriving at the resource after the specified fixed lead times. Releases in periods 2 and 3 contribute to the work input in period 3 at the work center performing operation j corresponding to the first and the third term

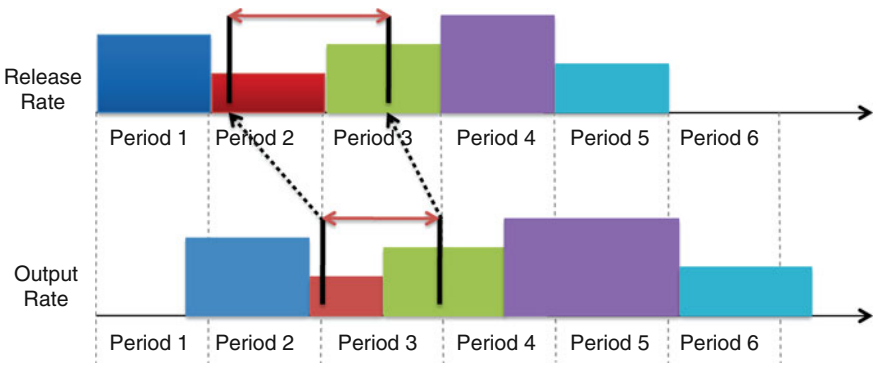

Fig. 16.3 Relationship between releases and loading with time-dependent lead times

in (16.11); the second term is not relevant here because the release interval Q only spans the two periods 2 and 3. Note that the lead times are associated with the boundary points between periods at the work center (not the boundary points between the release periods), and hence the lead time at the start of a period may not be the same as that at the end. The coloring indicates the correspondence between the releases and the arrival of the material at the resource.

Note that with this approach, we can write the output of product i in period t, denoted Y_{it}, as

$$Y_{it} = \sum_{\tau=1}^{T} R_{i\tau} w_i \tau_t \tag{16.12}$$

where the $w_{i\tau t}$ values denote the fraction of releases in period τ that contribute to output in period t. This results in a linear constraint. Note that if we were able to obtain the $w_{i\tau t}$ values correctly, we would no longer need an explicit capacity constraint of the form

$$\sum_{i=1}^{N} a_{ij} Y_{it} \leq C_t \tag{16.13}$$

since the weights $w_{it\tau}$ would reflect the ability of the resource to produce output over time. However, queuing theory tells us that these weights depend on the resource utilizations, which are determined by both the WIP profile in the system and the releases that are determined by the planning model in use.

An obvious solution to this dilemma is to embed the LP models using these lead times in an iterative scheme where the releases obtained from the solution to the production planning models are fed into a simulation model of the production facility to evaluate the realized lead times they would impose on the production system. Such approaches have been suggested in the literature, which we shall discuss below.

16.4.4 Iterative Approaches

The formulations described above based on workload-independent lead time distributions suggest an iteration scheme where an initial set of lead time estimates is used to create a plan, and the flow times that will be realized by the execution of the plan are predicted using a simulation or queuing model. These predictions of the realized flow times are then used to generate a new set of lead time estimates, with the procedure continuing until the change in lead time estimates from one iteration to the next is within some specified tolerance.

Specifically, recall that the output of product i in period t is estimated as

$$Y_{it} = \sum_{\tau=1}^{T} R_{i\tau} w_{i\tau t} \tag{16.14}$$

where the $w_{i\tau t}$ values denote the fraction of releases in period τ that contribute to output in period t and are computed based on the lead time estimates L_{it}. At the k'th iteration of the procedure, lead time estimates L_{it}^k are used to compute weights $w_{i\tau t}^k$. The execution of the resulting plan is then simulated to obtain a new set of lead time estimates, and the procedure continues until convergence. It is worth noting in passing that similar iterative techniques have been used in the job shop scheduling literature to develop dispatching rules that consider the remaining time until completion of a job at an intermediate stage of processing (Vepsalainen and Morton 1987, 1988; Lu et al. 1994).

Hung and Leachman (1996) tested this iterative scheme using the workload-independent lead time distribution described above and a simulation model of a wafer fabrication facility. They examined the rate of convergence of the flow time estimates to the actual flow time values in the simulation, and found that convergence to the correct expected flow time values can be quite rapid, but that the procedure can fail to converge in some cases which are not fully understood. Irdem et al. (2008) present an experimental study of the convergence of the method, which suggests that for production systems at high utilization levels it can be difficult to confirm convergence. This approach has the advantage of combining two off-the-shelf modeling techniques, linear programming and simulation, that practitioners are likely to be familiar with, in an iterative scheme that addresses the complex interdependency of releases and lead times. However, the need for a simulation model of the facility being planned requires both large amounts of data to construct and validate, and also increases run time significantly. The authors discuss several ways to reduce the computational burden of the simulation by focusing on highly utilized work centers. Hung and Hou (2001) substitute an analytical flow time prediction model for the simulation in the iterative scheme, and report results that compare favorably in convergence performance with those of the scheme using the simulation model. Byrne and Bakir (1999) use an iterative technique (iterations between the production planning model and a simulation model) to determine realistic estimates for the available capacities, Byrne and Hossain (2005) present an extended production planning model within this framework. Kim and Kim (2001) simultaneously update flow times and available capacities using this iteration scheme.

A rather different iterative technique has been proposed by Riaño(Riaño 2003; Riaño et al. 2003), where the $w_{i\tau t}$ values are estimated using a model of the transient behavior of a queuing network. In order to present the basic idea of the approach, we shall consider its application to a work center consisting of a single server; the extension of this model to multiple stages and servers is discussed in Riaño et al. (2006).

Riaño's approach is to consider the system from a queuing perspective. A job released into the production system at time s will see $Q(s)$ jobs ahead of it in the queue or in process. Hence the flow time of that job will be given by

$$W(s) = \sum_{k=2}^{Q(s)} S_k + S_1, \tag{16.15}$$

where $S_k, k = 2, \ldots, Q(s)$ denote the processing times of jobs ahead of this job in the queue, and S_1 the residual (remaining) processing time of the job currently in process. The distribution function of the flow time of the job introduced into the system at time s is then given by

$$G(s,t) = \sum_{n=0}^{\infty} F_1 * F^{n*}(t) P\{Q(s) = n\}, \tag{16.16}$$

where F_1 denotes the distribution function of the residual processing time of the job currently in process, "$*$" denotes convolution, and F^{n*} the n-fold convolution of the processing time distribution F at the server. Note that $G(s,t)$ describes a state-dependent flow time distribution that depends on the number of jobs $Q(s)$ in the system at the time s the job was released. We wish to develop an approximation to this function that will allow us to calculate approximate values of the weights w_{st} that relate the input in the s-th interval with the cumulative output by time t. These weights can be used to estimate the output of the resource under a particular release pattern. To develop such an approximation, the author assumes that this time-dependent lead time distribution will have the same form as the steady-state distribution function of the waiting time for an $M/G/1$ queue, which is given by

$$(1-\rho) \sum_{n=0}^{\infty} \rho^n F_e^{n*}(t), \tag{16.17}$$

where F_e is the equilibrium residual processing time distribution, derived assuming that the time a new job enters the system is uniformly distributed over the service time. This suggests an approximation of the form

$$G(s,t) = F_1 * (1-\beta(s)) \sum_{n=0}^{\infty} \beta(s)^n F_e^{n*}(t), \tag{16.18}$$

where $\beta(s)$ denotes a time-dependent traffic intensity. Noting that if we assume the service time distribution to be phase-type, then $G(s,t)$ will also be phase type, the author suggests heuristic estimates of $\beta(s)$, obtaining an approximation for $G(s,t)$ that depends only on the expected WIP level at time s, denoted by $\phi(s)$ and its time derivative $\phi'(s)$. Hence, to make the approximation to $G(s,t)$ we now need a viable technique for estimating $\phi(s)$ and $\phi'(s)$. This clearly depends on the pattern of releases into the production system, and so a recursive technique is used. Given a release pattern, we can compute estimates of $\phi(t)$ for every planning period t in a recursive manner, starting from period $t = 1$ and moving forward in time. If the processing time distribution at the server is phase-type (see, e.g., Neuts 1981), the author shows that these computations can be performed in an efficient manner. The resulting approximation to $G(s,t)$ yields approximate values of the w_{st}, which now correspond to the probability that a job released in period s will complete in period t. The author suggests a successive approximation method to compute the

weights, where for a given release pattern the weights are first estimated and then a planning problem is solved to estimate WIP levels. These new WIP levels are used to estimate new weights until the estimates of weights converge.

The larger pattern of the iteration is now clear: we begin with an initial release pattern and calculate initial estimates of the w_{st}. We then calculate a new release pattern using these weights, and repeat until, hopefully, convergence is achieved. As with the approach of Hung and Leachman (1996), the convergence behavior of this procedure is not well understood; when it converges, it converges quite rapidly but in other cases it can cycle through a limited number of solutions. Further experimental and theoretical work is necessary to understand this convergence issue, but the overall approach stands as a very interesting and novel approach to modeling workload-dependent lead times in production planning, with a strong theoretical underpinning. Interesting discussions in this direction are given by Hackman (2008).

16.4.5 Workload-Dependent Lead Time Models

A major disadvantage of the approaches with fixed lead times or lead time distributions is their failure to consider the relationship between resource utilization and lead times. The iterative approaches described above attempt to restore this relationship by using a simulation or queuing model of the plant to estimate the effects of the planned releases on the lead times, but the underlying optimization formulation retains this basic flaw. A number of authors have developed models that allow lead times to vary according to the resource utilization, where the models include some mechanism that selects an appropriate lead time for each planning period based on the resource loading in that period.

An interesting area that is closely related to production planning but has not been widely explored in this context is that of dynamic traffic assignment models (Peeta and Ziliaskopoulos 2001). The objective of these models is to manage the routing of vehicles through a road network in order to optimize some measure of performance. Clearly, individual traffic links (road segments) are subject to congestion, and hence considerable effort has been devoted to developing formulations that capture this characteristic, as well as mathematical models of the relationship between the volume of flow on a traffic link and the velocity of that flow.

One way to model congestion in traffic links is *time–space links* (Carey and Subrahmanian 2000). If two nodes i and j of a traffic network are connected by a spatial link (e.g., a road segment), this two-node-network can be expanded over time, which yields a network of time–space nodes (see Fig. 16.4). A time–space link is a connection between time–space nodes. The flow through a time–space link represents the numbers of vehicles that pass the nodes at the respective times and hence require the respective traversal time.

If congestion is modeled as a link traversal time that increases with flow through the link, this can be represented as capacities of the time–space links leaving node i at time t (maximum flow on time–space link (i,t)) that depend on the flow through node i at time t, i.e., the inflow to the time–space links leaving node (i,t).

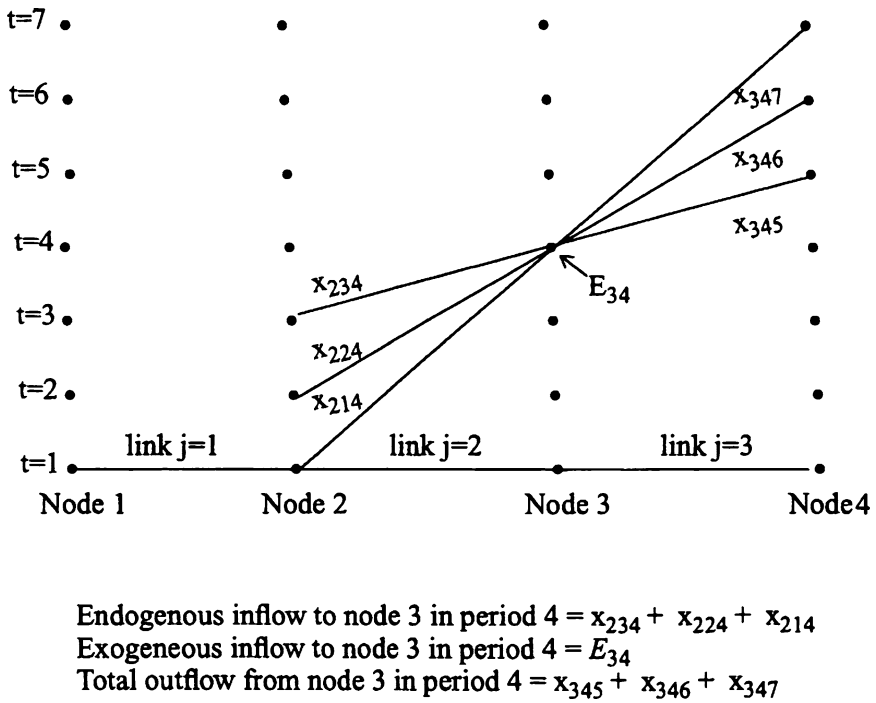

Fig. 16.4 Conservation of flows $x_{ji\tau}$ on a time expanded network (Carey and Subrahmanian 2000)

In the model of Carey and Subrahmanian (2000) the capacities of two time–space links leaving node (i, t) are usually greater than zero and the other time–space links are closed for the given inflow. As the inflow increases, the time–space links with positive capacities move to higher traversal times, implying a flow-dependent traversal time distribution that is stationary over time for a given inflow. The traversal time through a spatial link can be considered analogous to the flow time at a work center, and models with similar structure have been developed for order release planning in manufacturing.

A similar approach is described in Lautenschläger (1999). In order to consider load-dependent lead times for master production scheduling, the model determines the fraction of the planned production in a period t that has to be started one period ahead, in period $(t - 1)$. This fraction is a function of the planned utilization. Thus production on a resource can be performed in two modes, one with lead time of zero periods and the other with lead time of one period. The maximum production volumes that can take place in each mode are limited, which leads to a utilization-dependent lead time distribution.

Another related model has been proposed by Voss and Woodruff (2003). These authors assume a steady-state relationship between the utilization of a resource and the expected lead time at that resource. The basic idea is to discretize this curve and

use integer variables to construct constraints that ensure that only one segment of the discretized curve is active in a given time period.

In order to implement this formulation, the relationship between utilization and expected lead time is evaluated at discrete utilization levels BP_r, $r = 1, \ldots, R$. Let L_r denote the expected lead time value associated with the r'th utilization level BP_r, i.e., the expected lead time of the resource is assumed to be L_r while the utilization level is between BP_r and BP_{r-1}. The authors suggest setting the breakpoints BP_r such that each lead time L_r corresponds to an integer number of periods. If we now define the fraction of the available capacity of the resource required for one unit of product j, $j = 1, \ldots, P$, as a_j, and R_{jt} to be the amount of product j released in period t, the workload (utilization) of the resource in period t is given by

$$\rho_t = \sum_{j=1}^{P} a_j R_{jt}. \tag{16.19}$$

We now define binary variables y_{tr} that select a particular lead time value L_r to be active in a given period t as follows:

$$L_t = \sum_{r=1}^{R} y_{tr} L_r, \quad \text{for all } t, \tag{16.20}$$

$$\sum_{r=1}^{R} y_{tr} = 1, \quad \text{for all } t. \tag{16.21}$$

Additional constraints of the form

$$\sum_{r=1}^{R} \mathrm{BP}_r y_{tr} \geq \sum_{j=1}^{P} a_j R_{jt}, \quad \text{for all } t \tag{16.22}$$

are required to ensure that the lead time selected is consistent with workload. In addition, for any given period t, we require $L_t - L_{t+1} \leq 1$, giving

$$\sum_{r=1}^{R} y_{t,r} L_r - \sum_{r=1}^{R} y_{t+1,r} L_r \leq 1, \quad \text{for all } t. \tag{16.23}$$

This latter constraint is interesting in that it restricts the decrease in lead time from one period to the next to at most one period to avoid overtaking, i.e., material released into the system at a later time emerging before material released earlier. Similar difficulties arise in dynamic traffic assignment problems (e.g., Carey and Subrahmanian 2000).

To complete the formulation, the authors present an objective function that includes an explicit holding cost for WIP, based on Little's Law (see, e.g., Hopp and Spearman 2001), leading to the term

$$\sum_{\mathrm{t}=1}^{\mathrm{T}} \sum_{\mathrm{j}=1}^{\mathrm{P}} h_{jt} \sum_{\mathrm{r}=1}^{\mathrm{R}} y_{\mathrm{tr}} L_r R_{jt}. \tag{16.24}$$

This objective function is now nonlinear due to the product of the y_{tr} and R_{jt}, leading to a formulation that is computationally hard to solve.

16.4.6 Discussion of Models Based on Lead Time Distributions

While these models address the load-dependent nature of lead times directly, there are several shortcomings of these models:

- All the models described above assume the existence of a well-defined relationship between the workload or utilization of a resource and the expected lead time of that resource in that period. However, given that the planning models assume discrete planning periods of a fixed length and that the releases of work into the resource are varying over time, it is quite possible that the lead time incurred by work released in a given period may deviate quite substantially from that suggested by a long-run steady-state average relationship. The work of Riaño (2003) is a significant exception, explicitly addressing the transient nature of the queues involved, and thus merits further study.
- If the amount of work released decreases sharply from period t to period $t + 1$, the estimated lead time for the orders can decrease by more than one period from t to $t + 1$, implying overtaking (Voss and Woodruff 2003, p. 165; Carey and Subrahmanian 2000). This is unlikely to occur in practice and indicates that the models can lead to unrealistic results. Voss and Woodruff (2003) add a constraint that keeps the lead time from decreasing by more than one time bucket from t to $t + 1$, but this excludes decision alternatives and is not satisfactory from a theoretical point of view.

A number of researchers have proposed alternative approaches to these problems by developing formulations that do not consider the relationship between lead times and resource utilization explicitly, but instead use a relationship between the expected WIP level of a resource and its expected output in a given planning period. These clearing function based models will be discussed in the next section.

16.5 WIP-Based Models

The models discussed in the previous section all approach the problem of modeling the behavior of the production resource by computing a distribution of the lead times, a relationship between the time the work is released into the facility and the time it becomes available for consumption by the next stage. The distinguishing feature of this approach is the presence of a set of constraints implementing a lead time distribution used across different time periods, specifically in the material balance constraints. We now turn our attention to models where the lead time behavior

of the production resources is not represented in the balance constraints, but by introducing nonlinear terms in the constraints or the objective function. The former class of models introduce nonlinear constraints (that may be linearized for computational purposes) representing the relationship between some measure of the expected WIP level (including jobs in queue and in process) in front of the production resource in question over a planning period and the expected output of the resource over the planning period. These formulations, which we shall term *Clearing Function* models, are discussed extensively in the next section, and generally result in models with linear objective functions and convex nonlinear constraint sets. The latter, which we shall term *WIP* Cost formulations, use queuing analysis to develop an expression for the expected WIP holding cost which is then added to the objective function of the model. These formulations, by contrast, tend to yield models with convex nonlinear objective functions.

A fundamental difference between these models and those discussed up to this point is their explicit representation of WIP. In the LP formulations discussed until this point, although it is possible to recapture planned WIP levels as the difference between cumulative releases and cumulative output, the WIP level has no effect on the behavior of the production resource. The exception to this is the transient queuing-based approach of Riaño (Riaño 2003; Riaño et al. 2003) and the respective simulation-based approaches, but even in this case the WIP enters the formulation only through its role in determining the lead time distribution. The WIP-based models, on the other hand, explicitly represent the WIP level in front of the resource in a planning period with distinct material balance equations, in addition to the balance equations for finished goods inventory (FGI) included in the conventional LP formulations.

We believe that this explicit distinction between WIP and FGI leads to substantially richer models, since the two types of inventory serve different purposes and are controlled in different ways. If we anticipate a seasonal surge in demand in the future, our production planning model must ensure that we have sufficient FGI in place in time to meet the demand. Thus, FGI levels can be planned, and are an output of the planning process. On the other hand, the WIP levels are a consequence of our release decisions, and determine to a great extent the flow time and throughput performance of the production resource. WIP levels must be managed to ensure timely production at minimum cost, while FGI must be planned to ensure effective satisfaction of demand at minimum cost. Thus the explicit separation and modeling of these two different types of inventory offers the potential for significantly richer production planning models. The separation of WIP and FGI also has implications for developing production planning models that consider uncertainties, since in practice some of the functions of safety stocks can be assumed by WIP; we shall discuss this aspect in more detail in Sect. 16.8.

The models described in this section explicitly represent the flow of WIP through a production unit. They differ from the previous models in that they relate the output from a resource in a given planning period to some measure of the WIP level at the resource during the period. Thus the nonlinear relationships between decision variables arising from the presence of congestion are represented in the constraints

rather than in the objective function. We first describe the generic structure common to most models of this type and then concentrate on the technique used to incorporate the relationship between WIP, flow time, and output.

In models of this type, work centers are represented explicitly, and the material flow from order release through the required work centers and to the inventory of the final products or SKU's is represented by inventory balance equations. As in all models considered until now, material is modeled as a continuous medium, comparable to a fluid. Thus the approach is similar to fluid approximation in queuing theory (Kleinrock 1976, p. 60; Chen and Mandelbaum 1991) and to fluid relaxation in job-shop scheduling (Bertsimas and Sethuraman 2002; note that these fluid relaxation models are in continuous time and are usually deterministic). The planning horizon is again divided into discrete planning periods.

A variety of authors have discussed the relationship between system throughput and WIP levels. This is usually in the context of queuing analysis, where the quantities being studied are the long-run steady-state expected throughput rate and WIP levels. An example of this work is that of Agnew (1976), who studies this type of behavior in the context of optimal control policies. Spearman (1991) presents an analytic congestion model that describes a clearing function for closed production systems with processing time distributions with increasing failure rates. Standard texts on queuing models of manufacturing systems, such as Buzacott and Shanthikumar (1993), can be used to derive the clearing function – if not analytically, then at least numerically. Hopp and Spearman (2001) provide a number of illustrations of clearing functions for a variety of systems; for example, the relationship between WIP and throughput given in their practical worst case analysis represents a particular type of clearing function. Srinivasan et al. (1988) derive the clearing function for a closed queuing network with a product form solution. It is interesting to note that the concept of the clearing function has much in common with concepts developed in the dynamic traffic assignment literature. One such concept is that of the exit function, which defines the output of a traffic link in a time period as a function of the amount of traffic on the link at the start of the period (Merchant and Nemhauser 1978a, b; Carey 1987).

While the basic concept of a clearing function is quite intuitive, defining and implementing this concept in a rigorous and theoretically consistent manner is subject to some quite subtle difficulties that are not straightforward to resolve. Hence we will first discuss how optimization models of production planning problems may be formulated using clearing functions, assuming a valid clearing function can be generated. We shall then discuss the issues involved in estimating the clearing functions themselves.

16.5.1 Clearing Function Formulations

In its usual form the clearing function yields the expected aggregate output of a work center (e.g., hours of work, aggregated over the products) as a function of a suitable

measure of WIP, aggregated over the products. This WIP measure can be the average WIP of the period, the average WIP over a longer time (e.g., two periods), or the total available work of the period (termed *load*, defined as initial WIP plus input during the period). It is commonly assumed, based on both theoretical and empirical arguments discussed in the next section, that the clearing function is a concave nondecreasing function of the WIP measure used to define it. Our discussion in this section will closely follow the development by Asmundsson et al. (2009).

To illustrate this approach, we begin by extending the single-product single-stage formulation of Karmarkar (1989) to multiple products, which can be stated using the following notation:

X_{it} = number of units of item i produced in period t,
R_{it} = number of units of item i released into the stage at the beginning of period t,
W_{it} = number of units of item i in WIP inventory at the end of period t,
$\hat{W}_{it}$ = WIP measure used for the clearing function and defined in separate constraint (e.g., $\hat{W}_{it} = W_{i,t-1} + R_{it}$),
I_{it} = number of units of item i in finished goods inventory (FGI) at the end of period t,
ξ_{it} = time required to produce one unit of item i at the resource.

Let $f_t(\hat{W})$ denote the clearing function that represents the resource in period t, with W denoting the WIP measured in units of time (i.e., $W_t = \sum_i \xi_{it}\hat{W}_{it}$), and D_{it} the demand for item i (in units) in period t. Then a naive extension of Karmarkar (1989)'s single-product formulation to multiple products is

$$\min \sum_t (\phi_{it} X_{it} + \omega_{it} W_{it} + \pi_{it} I_{it} + \rho_{it} R_{it}), \tag{16.25}$$

subject to

$$W_{it} = W_{i,t-1} - X_{it} + R_{it}, \quad \text{for all } i, t, \tag{16.26}$$

$$I_{it} = I_{i,t-1} + X_{it} - D_{it}, \quad \text{for all } i, t, \tag{16.27}$$

$$\sum_i \xi_{it} X_{it} \le f_t\left(\sum_i \xi_{it}\hat{W}_{it}\right), \quad \text{for all } t, \tag{16.28}$$

$$X_{it}, W_{it}, I_{it}, R_{it} \ge 0, \quad \text{for all } i, t, \tag{16.29}$$

where $\phi_{it}, \omega_{it}, \pi_{it}$, and ρ_{it} denote the unit cost coefficients of production, WIP holding, finished goods inventory holding, and releases (raw materials) respectively, and ξ_{it} the amount of the resource (machine time) required to produce one unit of product i in period t. Note that the argument of the clearing function in constraint (16.28) is the total WIP level over all products i expressed in units of time (or, equivalently, workload). The first two sets of constraints enforce flow conservation for WIP and FGI separately. Since the formulation distinguishes between WIP and FGI, flow conservation constraints are required for both. Constraints (16.28)

represent the capacity constraint. Another interesting characteristic is that lead times do not appear in the formulation; they are represented implicitly by the nonlinear capacity constraints (16.28).

While this formulation appears intuitive, it can create significant modeling problems when applied in multiple product environments. Consider a situation where the system produces two products A and B, which consume capacity at the production resource in different amounts. The capacity constraint can be expressed as $X_A + X_B \leq f(\hat{W}_A + \hat{W}_B)$.

A solution with

$$X_A > 0, X_B = 0, \hat{W}_A = 0, \hat{W}_B > 0$$

may exist, despite the fact there is no WIP in the system that can be converted into finished product A. Hence the optimal solution to this formulation can maintain high WIP levels of the product for which it is cheapest to do so, using the capacity generated by this device (i.e., the high value of the clearing function attained by holding high WIP of the cheap product) to hold very low or no WIP of all other products. An alternative way of expressing this difficulty is that there is no link between the mix of WIP available in the period and the output mix during the period.

Asmundsson and his coauthors (Asmundsson et al. 2006, 2009) propose an approach in which the no-passing requirement is enforced on average rather than at the level of individual jobs. To do this, they assume that the mix of output will reflect the mix of WIP. This is equivalent to assuming a service discipline at the queue representing the production resource where no product is given priority over another. After some analysis described in detail in Asmundsson et al. (2009) this yields the following formulation:

$$\min \sum_t (\phi_{it} X_{it} + \omega_{it} W_{it} + \pi_{it} I_{it} + \rho_{it} R_{it}), \tag{16.30}$$

subject to

$$W_{it} = W_{i,t-1} - X_{it} + R_{it}, \quad \text{for all } i, t, \tag{16.31}$$

$$I_{it} = I_{i,t-1} + X_{it} - D_{it}, \quad \text{for all } i, t, \tag{16.32}$$

$$\xi_{it} X_{it} \leq Z_{it} f_t \left(\frac{\xi_{it} \hat{W}_{it}}{Z_{it}} \right), \quad \text{for all } i \text{ and } t, \tag{16.33}$$

$$\sum_i Z_{it} = 1, \quad \text{for all } t, \tag{16.34}$$

$$X_{it}, \mathrm{W}_{it}, I_{it}, R_{it}, Z_{it} \geq 0, \quad \text{for all } i, t. \tag{16.35}$$

This formulation will be referred to as the *Allocated Clearing Function* (ACF) model. The Z_{it} variables denote the fraction of the maximum possible output defined by the clearing function allocated to product i in period t. The intuition is that we wish to obtain a constraint that links the production of a given product to the

WIP level of that product alone, but the clearing function is defined in terms of the total WIP at the resource. This is, of course, what causes the difficulty described above: a high total WIP may result in a high maximum output level for the resource in a period, and the model may allocate this output in a manner that violates continuity of material flow. Assuming that all products at a resource will see the same expected lead time allows us to estimate the total WIP at the resource by the expression given as the argument of the clearing function in (16.33). The Z_{it} variables thus serve the dual purposes of scaling up the WIP for product i inside the parentheses of (16.33) to obtain a surrogate for total WIP on which the clearing function can operate, and then computing a fractional capacity for product i by multiplying the results. Asmundsson et al. (2009) prove that the total production of individual products will not exceed that suggested by the aggregate clearing function as suggested by (16.28).

The above formulation is a convex nonlinear program, due to the concave nature of the clearing function on the right hand side of constraints (16.33). However, an interesting and useful consequence of the partitioned formulation above arises when the concave clearing function is approximated using outer linearization. Since we assume the clearing functions are concave, they can be approximated by the convex hull of a set of affine functions of the form

$$\alpha^c \sum_i \xi_{it} \hat{W}_{it} + \beta^c \tag{16.36}$$

as

$$f(\hat{W}_t) = \min_c \{\alpha^c \hat{W}_t + \beta^c\}. \tag{16.37}$$

The $c = 1, \ldots, C$ index represents the individual line segments used in the approximation. In order to represent the concave clearing functions appropriately, we shall assume that the slopes of the line segments are monotonically decreasing, i.e.,

$$\alpha^1 > \alpha^2 > \ldots > \alpha^c = 0. \tag{16.38}$$

The slope of the last segment is set to zero to indicate that the maximum throughput capacity of the node has been reached, and adding WIP cannot increase throughput any further. To ensure that production cannot take place without some WIP being present, we impose the condition $\beta_1 = 0$ to ensure that the first line segment will pass through the origin. The capacity constraint in the CF formulation can now be replaced by the set of linear inequalities

$$\sum_i \xi_{it} X_{it} \leq \alpha^c \hat{W}_t + \beta^c, \quad \text{for all } c \text{ and } t. \tag{16.39}$$

Using this outer linearization to approximate the PCF model yields the following LP:

$$\min \sum_t \sum_i (\phi_{it} X_{it} + \omega_{it} W_{it} + h_{it} I_{it} + \rho_{it} R_{it}), \tag{16.40}$$

subject to

$$W_{it} = W_{i,t-1} - X_{it} + R_{it}, \quad \text{for all } i \text{ and } t, \tag{16.41}$$

$$I_{it} = I_{i,t-1} + X_{it} - D_{it}, \quad \text{for all } i \text{ and } t, \tag{16.42}$$

$$\xi_{it} X_{it} \leq \alpha^c \xi_{it} \hat{W}_{it} + Z_{it} \beta^c, \quad \text{for all } i, t, \text{ and }, c, \tag{16.43}$$

$$\sum_i Z_{it} = 1, \quad \text{for all } t, \tag{16.44}$$

$$Z_{it}, X_{it}, W_{it}, I_{it} \geq 0, \quad \text{for all } i \text{ and } t. \tag{16.45}$$

Notice that summing the set of constraints (16.43) over all i gives constraint (16.39), guaranteeing that the original constraint is satisfied. A consequence of the partitioning of the clearing function is that the clearing function constraint becomes linear, even with the Z-variables that were originally in the denominator since

$$Z_{it} f\left(\frac{\xi_{it}\hat{W}_{it}}{Z_{it}}\right) = Z_{it} \min_c \left\{\alpha^c \frac{\xi_{it}\hat{W}_{it}}{Z_{it}} + \beta^c\right\} = \min_c \{\alpha^c \xi_{it}\hat{W}_{it} + \beta^c Z_{it}\}. \tag{16.46}$$

In their experimental implementation, Asmundsson et al. (2006) assume the clearing function depends on the average WIP level over the time period, which they approximate as

$$\hat{W}_{it} = \frac{1}{2}(W_{i,t-1} + W_{it}).$$

16.5.2 Extensions to Multistage Systems

So far the clearing function formulations have been presented for a single-stage production system. In order to extend this to a multistage system, a number of extensions must be included. Since the model requires explicit modeling of WIP at each stage in order to compute the clearing function at each stage, we must represent the movement of material between stages; the output of one stage flows into the WIP of another. Another worthwhile enhancement is to distinguish between work centers that are potential bottlenecks and those that are unlikely to encounter significant congestion phenomena. The former can be represented explicitly using clearing functions, while the latter can be modeled using some form of workload independent delay, such as a fixed lead time, or a higher order delay as suggested by (Missbauer 2002a). While we assume in this model that groups of similar products are aggregated into product groups of families, the basic structure of the formulation remains the same if products are modeled individually, although, of course, a much larger formulation may result. We use the following notation:

Variables:

W_{jmt} – Work of product group j waiting at work center m at the end of period t.
I_{jt} – Finished goods inventory of product group j at the end of period t.
R_{jt} – Released work of product group j in period t.
X_{jmt} – Output of product group j from work center m in period t.

Note that for exposition we avoid the constants ξ of the above formulation and measure the variables in units of time (e.g., hours of work). Hence I_{jt} and R_{jt} are measured in hours of work, and W_{jmt} and X_{jmt} are measured in hours of work at work center m.
Parameters:

D_{jt} – Demand for product group j in period t (measured in hours of work).
C_{mt} – Capacity of work center m in period t.
$\tilde{p}_{jim}$– Average amount of work arriving at work center m when one unit of product group j is finished at work center i.
$z_{jim\tau}$ – Proportion of the output of product group j from work center i to work center m that arrives at m in τ periods after completion at i. These are not included to capture congestion-related lead times, but rather to represent delays such as shipping or flow through non-bottleneck work centers that are not subject to significant congestion effects, similar to those modeled by Hackman and Leachman (1989).

As for the single stage models, this formulation requires separate flow balance equations for WIP and FGI at each stage

$$W_{jmt} = W_{j,m,t-1} + \sum_{i=1}^{M} \sum_{\tau=0}^{\infty} X_{j,i,t-\tau} \tilde{p}_{jim} z_{jim\tau} + \sum_{\tau=0}^{\infty} R_{j,t-\tau} \tilde{p}_{j0m} z_{j0m\tau} - X_{jmt}, \quad \text{for all } j, m, t. \tag{16.47}$$

These constraints model the flow of WIP at the bottleneck work centers. The sources of input are the output of the other work centers and the release of orders. The work input can be delayed by transportation, non-bottlenecks (see below), etc. Work center $i = 0$ denotes the beginning of the line where order release takes place. The finished goods inventory is then represented as

$$I_{jt} = I_{j,t-1} + \sum_{m=1}^{M} \sum_{\tau=0}^{\infty} X_{j,m,t-\tau} \tilde{p}_{jm0} z_{jm0\tau} - D_{jt}, \quad \text{for all } j \text{ and } t. \tag{16.48}$$

Note that inflows into FGI can also be delayed after completing their last production operation. Work center index 0 in p and z denotes the completion of the product.

Constraints (16.47) and (16.48), together with an appropriate objective function, the clearing function-based capacity constraints (16.43) for the bottleneck work

centers and the partitioning constraints (16.44), constitute a complete clearing function-based formulation for a multiproduct, multistage production system. Asmundsson et al. (2006, 2009) give a slightly different formulation that is based on Hackman and Leachman (1989), where an additional set of decision variables is added to represent material transfer between stages; note the formulation above assumes that material transfer to the next stage begins directly upon completion of processing at the current stage.

The computational results obtained using the clearing function formulations are quite promising, although more experimentation is clearly needed to be able to draw strong conclusions. Probably the most complete studies of these formulations at present are those of Asmundsson et al. (2006, 2009). The later study examines the performance of clearing function formulations in simple serial production systems, and concludes that when the clearing function is correctly estimated the clearing function formulations produce production plans that are much more aligned with the ability of the plant to execute them compared to models with fixed lead times. This study describes a systematic procedure for estimating the clearing functions from simulation data and fitting to the functional form (16.58) below using a nonlinear optimization algorithm. Asmundsson et al. (2006) compare the performance of the clearing function formulations to that of a conventional LP model in a reentrant line derived from a semiconductor wafer fabrication facility, and find that even when a simple visual technique is used to fit the clearing functions to the data, the CF models yield significantly better on time delivery than the LP models with fixed lead times.

16.5.3 Estimation of Clearing Functions

The reader will have noted that until now we have described formulations that make use of the clearing function concept but have not discussed how the clearing functions are estimated. In this section we provide a more formal definition of clearing functions and discuss this very important issue.

A *clearing function*, introduced by Graves (1986), can be defined as either the expected or maximum output of a work center (a relatively homogeneous group of production resources scheduled as a unit) in period t as a function of some measure of WIP (e.g., average WIP or planned available work) in period t and the maximum capacity of the work center C_{it}. In the following discussion we shall use as the WIP measure the *load* of work center i in period t, denoted by A_{it}, defined as the WIP at the beginning of period t plus the planned work release R_{it} in period t[3]. Thus, the load (total available work) at work center i in period t is given by

[3] This assumes a production unit consisting of a single work center, which we assume for exposition. For a multistage system as in Sect. 16.5.2, the work release R_{it} is replaced by the work input from release and from the other work centers as in (16.47).

$$\Lambda_{it} = W_{i,t-1} + R_{it}. \tag{16.49}$$

The clearing function for work center i is then a functional relationship of the form

$$X_{it} = f_i(W_{i,t-1} + R_{it}; C_{it}) \tag{16.50}$$

Note that in conventional LP models, only the maximum capacity C_{it} would be considered in a capacity constraint.

This approach has been followed by the majority of researchers proposing models of this type (e.g., Karmarkar 1989; Missbauer 2002a; Selçuk et al. 2007). The clearing function models the impact of the fact that the *actual* load (available work) is a random variable at the time of planning and thus can be lower than the planned value, and work arriving during the period can arrive later than expected and thus cannot be processed during the period. Uzsoy and his coworkers (Asmundsson et al. 2006; Hwang and Uzsoy 2005) have adopted a slightly different approach where the clearing function is defined as a function of the expected, time–average WIP level during the planning period (see Sect. 16.5.1). While this difference in approaches requires some modifications to the details of the formulations and the procedures used to estimate the clearing functions, the basic structure of the approach remains the same.

Figure 16.5, derived from Karmarkar (1989), depicts several examples of clearing functions considered in the literature to date, where X denotes the expected throughput in a planning period. The "constant proportion" linear clearing function of Graves (1986) and Parrish (1987) allows unlimited output in a planning period, but ensures fixed lead time. This type of clearing function is not generally applicable to order release models, because it can yield capacity-infeasible output levels at high levels of WIP. Alternatively, it requires an assumption that the production rate of resources can be managed such that the fixed lead time is always maintained.

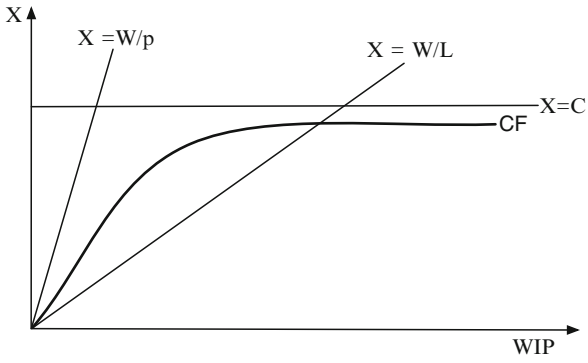

Fig. 16.5 Examples of clearing functions (from Karmarkar 1989)

However, by linking production rate to WIP level, it differs from the fixed delays used in most LP models, where the output of a production process is simply the input shifted forward in time by the fixed lead time. Orcun et al. (2006) illustrate the differences in the transient behavior of the production system under this and several other clearing function models.

The horizontal line $X = C$ corresponds to a fixed upper bound on output over the period, but without a lead-time constraint it implies that production can occur without any WIP in the system if work input and production are synchronized. This is reflected in the independence of output from the WIP level, which may constrain throughput to a level below the upper bound by starving the resource. This approach is implemented in, for example, the MRP-C approach of Tardif and Spearman (1997) and most LP approaches such as that of Hackman and Leachman (1989), but is supplemented with a fixed lead time as described in Sect. 16.3. The linear clearing function of Graves (1986) is represented by the $X = W/L$ line, which implies a lead time of L periods that is maintained independently of the WIP level. Note that if WIP and output are measured in the same time units (e.g., hours of work), the slope of the proportional part of the function is $1/L$, where L is the average lead time. However, as seen in the figure, this model may suggest infeasible output levels when WIP levels are high. If a fixed lead time is maintained up to a certain maximum output, we have the relationship $X = \min\{W/L, C\}$. When the parameters of the Graves clearing function are set such that the lead time is equal to the average processing time, with no queuing delays at all, we obtain the line $X = W/p$, where p denotes the average processing time. Assuming that lead time is equal to the average processing time up to a maximum output level gives the "Best Case" model $X = \min\{W/p, C\}$ described in Chap. 7 of Hopp and Spearman (2001). It is important to note that the workload-independent fixed lead time discussed in Sect. 16.4 differs from the linear model of Graves in that the former does not link output to WIP, while the latter does. Orcun et al. (2006) illustrate the differences between these models using system dynamics simulations.

It is apparent from the figure that the clearing function always lies below the $X = W/p$ and $X = C$ lines. For most capacitated production resources subject to congestion, limited capacity leads to a saturating (concave) shape of the clearing function. It is important to note that the nonlinear shape of the clearing function is not purely due to the presence of random variability in arrival and service processes at the production resource but can arise even in completely deterministic capacitated production systems, as shown in Karmarkar (1993). A number of recent papers (e.g., Asmundsson et al. 2009; Selçuk 2007) provide analytical support for the concave shape of the clearing function (see Sect. 16.5.3.1).

We now discuss techniques for estimating the clearing function associated with a set of production resources.

16.5.3.1 Analytical Approaches

A common approach to estimating clearing functions is to derive them using steady-state queuing analysis. It can be shown that for the $M/G/1$ model in steady state, the average throughput $E(X)$ is related to the expected WIP level $E(W)$ as follows:

$$E(X) = C \cdot \frac{E(W)}{E\,(W) + k} \tag{16.51}$$

where

$$k = \frac{\mu\sigma^2}{2} + \frac{1}{2\mu} \tag{16.52}$$

Here $1/\mu, \sigma^2$ denote the mean and variance of the processing time distribution and C the maximum capacity per period of the resource in hours of work. This is the same functional form as in Karmarkar (1989) (16.57 below), but (16.57) relates the output of the resource in period t to its load in period t, and the functional form of (16.57) is not supported by queuing models. (Missbauer 2002a) shows that for the $M/G/1$ model in equilibrium, the expected output $E[X_t]$ and expected load $E[\Lambda_t]$ of a work center are related as follows:

$$\begin{aligned} E[X_t] = \; & \frac{1}{2}\,(C + k + E[\Lambda_t] \\ & - \sqrt{C^2 + 2Ck + k^2 - 2\,C\;E[\Lambda_t] + 2k\;E[\Lambda_t] + E[\Lambda_t]^2}. \end{aligned} \tag{16.53}$$

The parameter k can be calculated analytically using (16.52) for the $M/G/1$ model, but can also be determined from empirical data or from simulation results.

For single-server work centers it seems reasonable to use (16.53) as a regression function with k and possibly also C as parameters. This need not be the case for multiple-server work centers, because in this case the first derivative of the average flow time with respect to the average WIP level is very low for low levels of WIP, where servers can be expected to be idle and the average waiting time is close to zero; this shape can be seen in Fig. 16.1). In this region the clearing function is nearly linear; the average flow time being insensitive to WIP implies a linear clearing function according to Little's Law (see Graves 1986 for the proof for the case of a discrete-time clearing function), which is not covered by (16.53). If the clearing function is approximated by piecewise linearization as discussed in the previous section, this is not a big problem.

The clearing function formulation (16.53) is derived from a steady-state model, which is a severe limitation. Selçuk (2007) derive a clearing function for a production resource with exponentially distributed service times, assuming the work available for the period is available at the time it is required for processing, without requiring steady state. They prove that this clearing function, which they refer to as a short-term nonlinear clearing function, is concave in the resource load defined above. Asmundsson et al. (2009) generalize this result. The general problem of determining a theoretically consistent clearing function for a single period without

steady-state assumption for stochastic arrival and departure processes is largely unsolved and an important research topic, which we discuss further in Sect. 16.5.4.

Based on this discussion, we will be interested in clearing functions of the form $X_{it} \leq f_i(\Lambda_{it})$ with the following properties:

$$f_i(\Lambda_{it}) \leq \Lambda_{it} \quad \text{for } \Lambda_{it} \geq 0 \tag{16.54}$$

$$\frac{\mathrm{d}f_i(\Lambda_{it})}{\mathrm{d}\Lambda_{it}} \geq 0, \quad \text{for} \Lambda_{it} \geq 0, \tag{16.55}$$

$$\lim_{\Lambda_{it} \to \infty} f_i(\Lambda_{it}) = C_{it}. \tag{16.56}$$

The use of nonlinear clearing functions requires consideration of the optimization technique used to obtain solutions to the resulting formulations. Piecewise linearization is frequently used (Missbauer 1998, 2002; Asmundsson et al. 2006). If the clearing function is nonlinear and concave, models of a reasonable size can be solved by nonlinear programming since they generally result in formulations with convex constraints and objective functions. Hwang and Uzsoy (2005) discuss a clearing function model of this type including lot sizing, while Srinivasan et al. (1988) discuss possible solution techniques for their model that involves nonlinear clearing functions. Continuous-time models that assume a deterministic flow at the work centers (fluid relaxation) can be solved either by exact methods (e.g., for the makespan objective, see Bertsimas and Sethuraman (2002) or heuristically (e.g., for the holding cost objective; see Bertsimas et al. (2003). Continuous-time models are not considered because the integration of clearing functions into these models has not yet been studied and remains a topic for future research.

Empirical work to date (e.g., Asmundsson et al. 2009) suggests that optimization models of aggregate material flow can be quite sensitive to the properties of the clearing function, especially when at high WIP levels the maximum output decreases. This is often encountered in modeling traffic flows, where it takes the form of a flow-density relation as seen in Fig. 16.6 that relates the density of the traffic on the link to the flow velocity of traffic through the link. In our production planning context, density corresponds to WIP and flow velocity to output rate. The chapter in this handbook by Armbruster and Lefeber presents a number of models where this type of approach is used to develop continuum models of flow through a manufacturing system in a manner analogous to that used to study traffic flows.

In manufacturing systems a decrease in output at high WIP levels can occur in two cases: (1) when workers work less efficiently under high pressure or (2) when long queues, and hence long average flow times, threaten due-date performance and require preemptive sequencing rules in order to pull ahead urgent orders reducing capacity due to additional setups. This can lead to solutions that must be considered infeasible if an appropriate arrival rate control policy is not applied. Van Ooijen (1996, p. 139 ff., especially pp. 144–146) describe the effect on system behavior, while Van Ooijen and Bertrand (2003) present an arrival rate control policy for this case. Haxholdt et al. (2003) demonstrate the possibility of oscillating behavior and chaos under more sophisticated assumptions on the arrival and departure

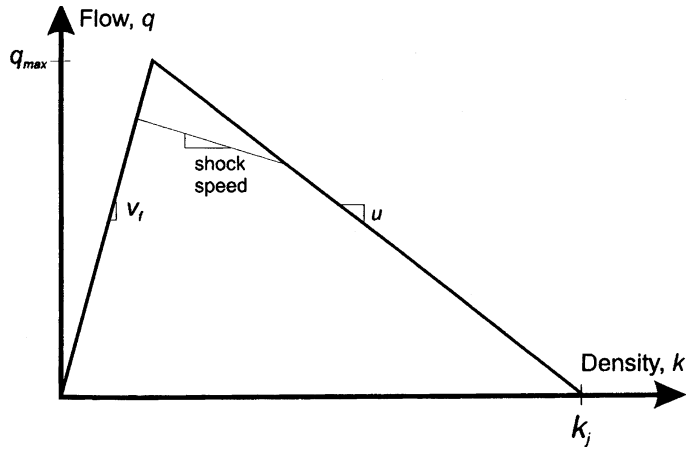

Fig. 16.6 Flow-density-relation (Cassidy 2003, p. 183)

process of queuing systems. The chapter by Elmaghraby in this volume also discusses possible shapes of clearing functions that may apply in this type of situation. Therefore, the shape of the clearing function has to be considered carefully in every case. In particular, a service rate that decreases if the level of WIP exceeds a certain threshold value can have extreme consequences on the system behavior, so this possibility has to be excluded or considered carefully when the clearing function is estimated. A literature review on the relationship between worker behavior and inventory level, especially in flow lines, can be found in Powell and Schultz (2004).

16.5.3.2 Empirical Estimation Techniques

A number of authors have suggested an empirical approach to estimating clearing functions, where a functional form with the desired properties is postulated, and then fit to data obtained either from an industrial facility or a simulation model using some form of regression analysis. Karmarkar (1989) uses the following functional form for the clearing function:

$$X_t = \min\left[C_t \frac{\Lambda_t}{\Lambda_t + k}; \Lambda_t\right], \tag{16.57}$$

where X_t denotes the output in period t, Λ_t the load of the resource at the start of period t, and C_t the maximum capacity of the resource available in period t. The shape parameter k is estimated by the user. The functional form of (16.53) above is an adaptation to models in discrete time. Srinivasan et al. (1988) suggest an alternative functional form

$$f(\Lambda_t) = C_t(1 - e^{-k\Lambda_t}) \tag{16.58}$$

where k is again a user-estimated shape parameter. Asmundsson et al. (2009) use this latter functional form, and give an extensive discussion of various issues in collecting simulation data for the purpose of fitting this type of clearing function. Asmundsson et al. (2006) use a visual fit of linear segments to simulation data to develop a clearing function formulation for a scaled-down semiconductor wafer fabrication facility with unreliable equipment and reentrant flows. There appears to be very little published literature using industrial data to fit clearing functions: the only paper we are aware of is Fine and Graves (1989), which motivated Graves' work on linear clearing functions.

If a saturating clearing function can be assumed, the estimation of the clearing function from empirical or simulated data (combinations of WIP or load and output for several periods) is essentially a curve-fitting procedure. The problem can be formulated as estimating the parameter values of a nonlinear function such as (16.58). If the clearing function is approximated by a set of N tangents, the parameters of the tangents can be derived from a nonlinear regression function (Missbauer 1998, p. 410 ff.), obtained directly from the observed data by numerical methods (Missbauer 1998, p. 407 f.) or by visual methods of curve fitting (Asmundsson et al. 2006). Estimating the clearing function can be difficult if the data include effects of machine downtimes. In this case, for some periods the average WIP is high and the output is low (because the work center has been down and the WIP could not be processed), and a simple curve fitting procedure would be misleading. Some sample data generated by simulation is shown in Fig. 16.7, which plots total throughput in a period against the average WIP in the period.

Asmundsson et al. (2006) describe a way to correctly estimate the clearing function in this case using multiple replications of simulation experiments. However, even in this case intuitive approaches can give poor results.

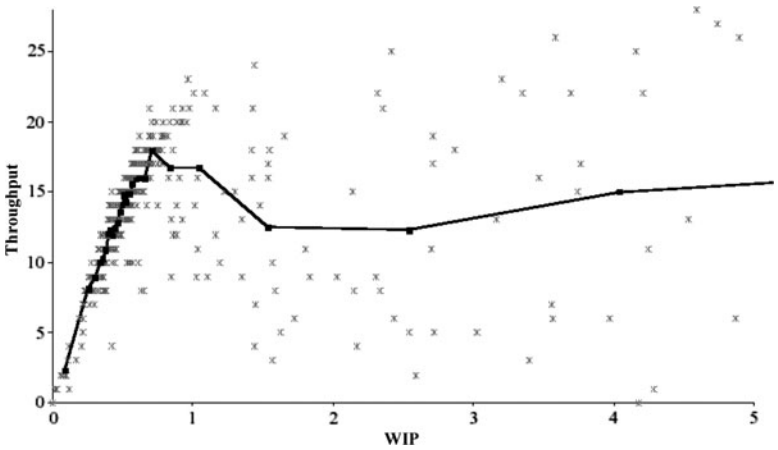

Fig. 16.7 Curve fitting in the case of machine downtimes – throughout vs. average WIP (Asmundsson et al. 2006)

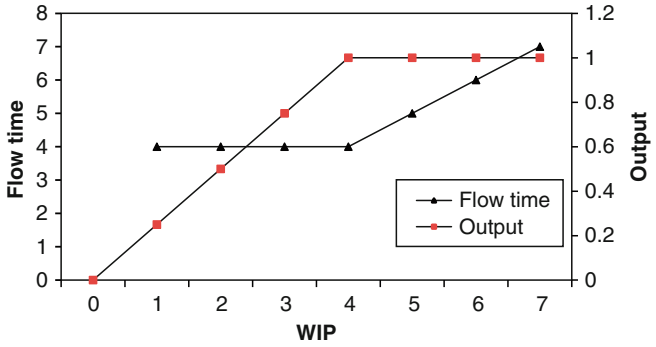

Fig. 16.8 Ideal clearing function for a paced flow line with four work centers (following Hopp and Spearman 2001, p. 221ff.)

An alternative way to determine the clearing function is the following: It is reasonable to assume that for very low WIP levels the clearing function is linear, and beyond a certain WIP level (that acts as buffer against variability) the work center can operate at full capacity. In an idealized situation, such as a deterministic paced flow line with equal processing times at each work center, the clearing function consists of these two parts (see Fig. 16.8).

In most practical cases the output will be lower for a certain range of WIP levels, and the problem of estimating the clearing function can be viewed as that of estimating the deviation of the clearing function from the ideal shape. Methods for this estimation are described in Nyhuis and Wiendahl (2003, p. 61 ff). Selçuk (2007) follows a similar approach and controls the shape of the clearing function by a parameter that reflects whether a more optimistic (overestimating throughput for a given WIP level) or conservative approach is applied (p. 115 f.).

The clearing function models described so far can become quite large if the number of work centers and product groups are large. If piecewise linearization of the clearing function is used, the number of linear constraints can be very high. If a reduction in model size is necessary, the non-bottleneck work centers can be eliminated from the model if a linear clearing function can be assumed for these work centers. In this case the non-bottlenecks can be represented as load-independent delay distributions (for the mathematical proof, see Missbauer 1998, p. 267 ff). This technique is similar to the delay functions in system dynamics models (Forrester 1962, p. 86 ff.).

16.5.4 Limitations of Clearing Function Models

It is evident from the last section that if the model of a clearing function as defined above is accepted, there are a number of open questions for future research. In this section, we examine the limitations of a clearing function that relates the expected

or maximum output of period t to the planned WIP or load in period t. As discussed above two different approaches have been used to estimate these functions. Empirical methods postulate a particular functional form having the "right" properties and fitting a curve of this form to data obtained from historical observations of the production system or from simulation. Other researchers have used steady-state queuing models to derive closed-form expressions for clearing functions. Both these approaches implicitly assume that the form of the clearing function can be treated as invariant over at least a range of system operating conditions; in empirical methods, over the range of environmental conditions represented in the data set used to fit the clearing function, while in steady-state queuing methods, over the entire life of the system. Another way of phrasing this problem is that both these approaches to estimating clearing functions produce a clearing function that maps the expected value of one random variable describing the WIP level (at the start of the period, or the expected WIP level over the period) to the expected value of another, the total output of the resource in a given period, and this relationship is assumed to be time-invariant.

The two different approaches suffer from different difficulties under this paradigm. When empirical methods are used, it is assumed that the clearing function that is obtained from empirical or simulated data as the estimated functional relationship between the *actual* WIP and output is a valid estimation of the functional relationship between the *planned* WIP and expected or maximum output in the context of the planning model (Missbauer, forthcoming). The possible problems resulting from this are still not well understood today.

Next, there is a classical sampling issue – we assume that the fitted clearing function will be able to represent the behavior of the system under conditions not encountered in the data sets from which the function was generated. In addition, experimental work to date has shown that fitting clearing functions to empirical data is by no means a straightforward exercise. Asmundsson et al. (2009) present a detailed procedure for estimating clearing functions from simulation output, and apply their procedure to a production system with significant machine failures. Their approach consists of three basic stages: collect the simulation data, fit a functional form to this data using least-squares regression, and then piecewise linearize the resulting concave function using a nonlinear optimization model to minimize the deviation of the linearized model from the original concave function. An example of the data obtained in this experiment is shown in Fig. 16.9 below.

The profusion of points on the WIP axis, denoting periods in which WIP was present but the machine was unable to produce output due to being down, results in a least-squares approach giving a poor fit. (Note that the least-squares approach yields the mean value of the output conditional on the WIP level; see Davidson and MacKinnon 1993, p. 41.) Inspection of this figure, specifically the point at which the fitted line reaches its maximum, suggests that the fitted function significantly underestimates the amount of WIP required for the resource to achieve its maximum output as approximately 50 units as opposed to a reality of about 400 units.

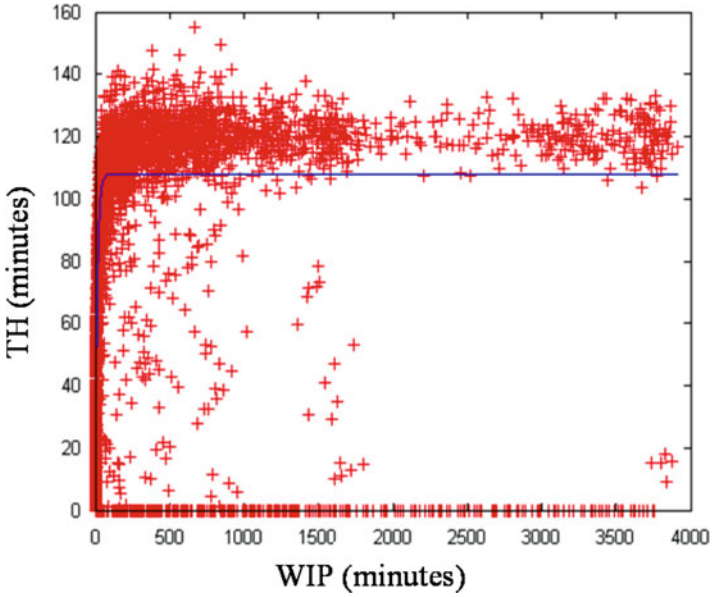

Fig. 16.9 CF data (WIP vs. TH) for a Work center from a Simulation Study

The effect of this error on the planning model is disastrous: it consistently releases too little material into the line too late, resulting in significant backlogs and missed demand.

In order to remedy this situation, the authors adopted an alternative fitting approach where they sought a curve such that a specified percentage of the data points fell above the curve, i.e., the fitted curve represents a percentile of the data, which corresponds to a quantile regression. The results of this approach for different percentile values are shown in Fig. 16.10. As the percentage of data points required to lie above the fitted curve increases, the fitted curve shifts to the right, providing a more realistic representation of how much WIP is required at this work center to achieve maximum output, based on the simulation results, and resulting in a planning model that yields better backlog results than an LP model. This heuristic approach proposed clearly needs a better theoretical justification, even though it works quite well in these experiments. These results highlight how a poorly fitted clearing function can result in poor performance of the planning models derived from it.

It is important to remember that the clearing function as defined in this paper is defined with relation to a planning period of a specific duration. When steady-state queuing models are used to derive an expression for the clearing function, we are assuming that the planning period is long enough that the behavior of the production resource being modeled is represented by a steady-state model to an acceptable degree of accuracy. However, in the production planning environment we are changing the releases, and therefore the workload, of the system in each planning period,

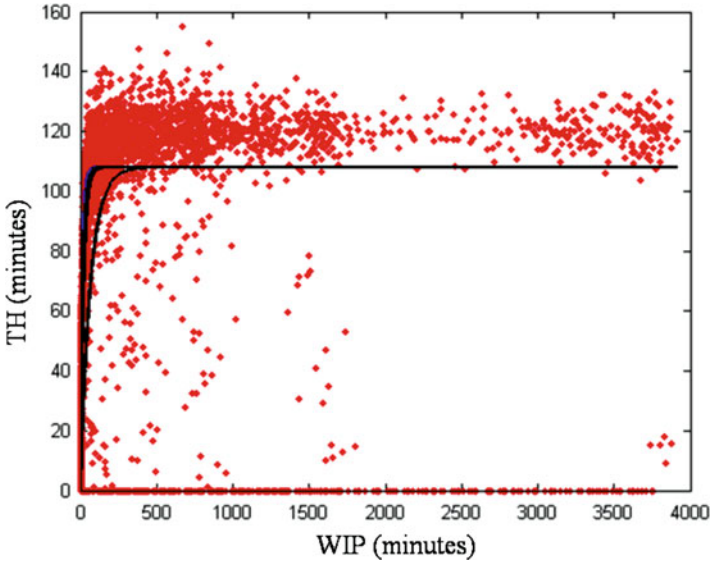

Fig. 16.10 Alternative percentile-based CF curves

calling into question whether the production resources ever attain the steady state required by conventional queuing analysis. We examine this question by comparing clearing functions derived for an $M/M/1$ queuing system in transient and steady state.

We consider an $M/M/1$ queuing system and define as a time unit the average operation time of an order, hence the service rate $\mu = 1$ without loss of generality. Hence the arrival rate λ is equal to the traffic intensity ρ. The length of a planning period is $\omega = 5$ time units. The clearing function for an arbitrary period t can be derived by recalling that the utilization u_t of the server in period t is the fraction of the period in which there is at least one order in the system and is given by

$$u_t = 1 - \frac{1}{\omega} \int_{\omega(t-1)}^{\omega \cdot t} p_0(\tau) d\tau \tag{16.59}$$

where $p_n(\tau)$ denotes the probability of n customers in the system at time τ, while $\omega(t-1)$ and $\omega \cdot t$ denote the beginning and the end of period t. The expected output $E[X_t]$ in period t is then

$$E[X_t] = u_t \cdot \omega \tag{16.60}$$

and the expected number of orders in the system at time τ, denoted $E\,[Ls(\tau)]$, is[4,5]

$$E\,[Ls(\tau)] = \sum_{n=0}^{\infty} n\; p_n(\tau). \tag{16.61}$$

This is also the average WIP, measured in units of time, at time τ, because the average service time of the orders in the queue is 1, and the expected remaining service time of the order at the server is also 1 due to the exponentially distributed service time. The expected load in period t, denoted by $E[\Lambda_t]$, is the average WIP at the beginning of period t plus the average input during period t:

$$E(\Lambda_t) = E\,[Ls(\omega(t-1))] + \rho\omega \tag{16.62}$$

The time-dependent state probabilities $p_n(\tau)$ for an $M/M/1$ system starting at the origin can then be calculated as (Stange 1964):

$$p_o(\tau) = 1 - \int_o^{\tau} \frac{e^{-(\rho+1)y}}{y} \cdot \sqrt{\rho}\,\text{Bessel}\; I_1(2\sqrt{\rho}y)\; dy \tag{16.63}$$

$$p_n(\tau) = \int_0^{\tau} \frac{e^{-(\rho+1)y}}{y} \left[n\rho^{\frac{n}{2}}\,\text{Bessel}\; I_n - (n+1)\rho^{\frac{n+1}{2}}\,\text{Bessel}\; I_{n+1}\right] dy, \quad \text{for } n > 0. \tag{16.64}$$

The expressions for the conditional probability of j customers in the system at time t given i customers in the system at time 0 can be found in Cohen (1969, p. 82 ff., p. 178).

The clearing function in period t can be calculated as a parametric curve with the average load in period t, $E[\Lambda_t]$ (16.61–16.64) on the x-axis and the expected output in period $tE[X_t]$ (16.59, 16.60, 16.63) on the y-axis. The arrival rate $\lambda = \rho$ is the control variable that yields the values for $E[\Lambda_t]$ and $E[X_t]$. The arrival rate is assumed to be the same from time $\tau = 0$ to the end of the period under consideration. Figure 16.11 shows the clearing functions for periods 1 and 2 and for a period after steady state has been reached. Note that in Fig. 16.11 the arrival process is stochastic for all work available. If work definitely will be available (e.g., at the beginning of the first planning period, where the actual WIP is known), the shape of the clearing function will be more extreme because the expected output cannot be lower than the minimum of the available work at the start of the period W_{t-1} (which we assume to be known) and available capacity. It can be shown (Missbauer, forthcoming) that

[4] Due to the computational complexity, the summation in (16.61) has been performed for $n = 0,\ldots, 80$ in the numerical examples below. This ignores at most 1.5% of the cases (for $\rho = 0.95$ in steady state), in most cases the error is close to zero.

[5] The index for the periods (discrete time) is denoted as subscript, the continuous time is denoted in parenthesis.

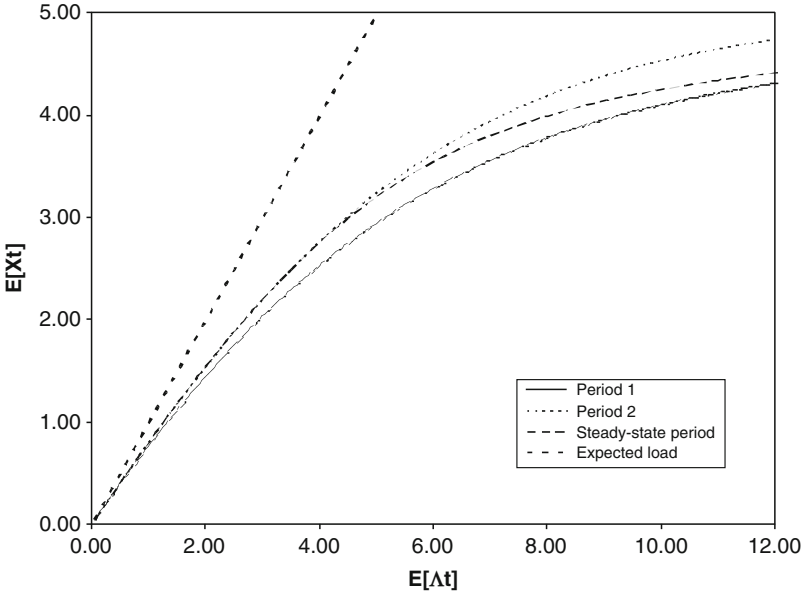

Fig. 16.11 Clearing functions for periods 1 and 2 and for a steady-state period of an *M/M/1* system starting at the origin. μ =1, length of a period $\omega = 5$. Note that identical values for $E[\Lambda_t]$ correspond to different values for the arrival rate λ for different periods

the shape of the clearing function (i.e., the functional relationship between expected load and expected output, with the expectation defined at the time of planning) depends on the variance of the initial WIP and of the planned input.

Figure 16.11 shows that if the system is not in steady state or in a specified transient phase there is no fixed functional relationship between expected load and expected output. The relationship changes with the phase of the transient state. This leads to a planning circularity that must be considered as a substantial problem of clearing function models. The estimated clearing function is based on assumptions about the dynamic behavior of the system. Hence, any order release plan derived using the clearing function can affect the dynamic behavior of the system, and hence the shape of the clearing function it is based on. This implies that *order release determines the validity of the assumption it is based on.* There is no evidence that the *assumed* shape of the clearing function is consistent with the *observed* shape(s) of the clearing function since the shape can change over time as seen in Fig. 16.11. We do not even know whether a consistent solution is possible, but it can be expected that this is not the case, because clearing function models assume that the clearing function is the same for each period, which need not be the case if transient/stationary phases occur during the planning horizon. Therefore the clearing function model can lead to systematic errors, and it is an empirical question whether or not the level of accuracy provided by the models is acceptable in practice. The experimental results of (Asmundsson et al. 2009) suggest that under some conditions

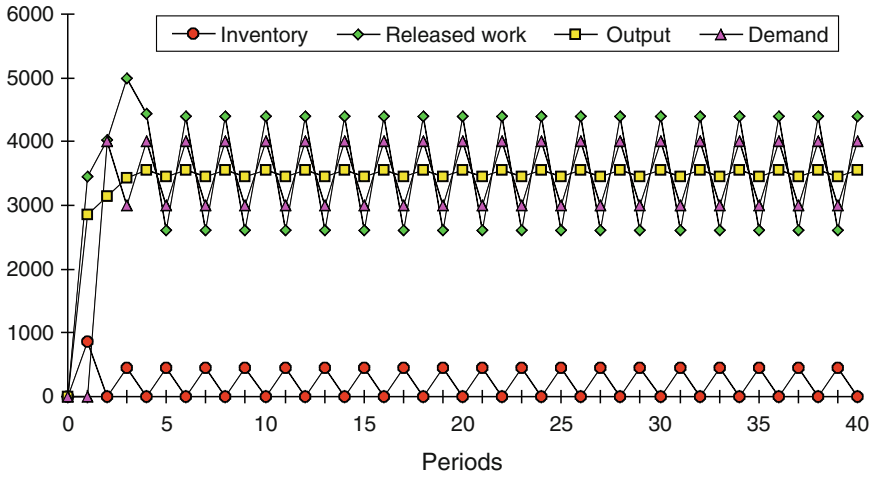

Fig. 16.12 Results of an optimization model for one work center and idealized underutilization

the performance of a planning model based on clearing functions can be very sensitive to how well the clearing function represents the actual system, which suggests caution in using these models until these issues are better understood.

Another limitation of clearing function models is that due to their use of discrete planning periods, they ignore any transient effects that arise at the boundaries between periods. While this limitation is shared with all planning models based on discrete time periods, it is worthy of note, and as far as we are aware has not been the subject of extensive research, as will be discussed further below.

Some characteristics of clearing function models can be seen from Fig. 16.12 which is an optimization result for one work center. The demand oscillates between 3,000 and 4,000 min of capacity per period, which is below the capacity (4,500 units/period). The variations in the planned output are much lower because of the nonlinear increase in the amount of WIP that is required. But the amount of released work exaggerates the demand variation. Figure 16.12 exhibits nervous behavior that has been reported as a property of optimization models if steady-state properties are assumed to hold for short periods (Lautenschläger 1999). Karmarkar (1993, p. 317), also states that "what happens in the transition between periods is not clear." In Fig. 16.12 it is difficult to decide whether this behavior is truly optimal, because this would require a model that incorporates the actual characteristics of the transient state.

Figure 16.13 exhibits the due-date deviations achieved by the clearing function model of Missbauer (2002a) for a highly utilized production unit with five bottleneck work centers. The average due-date deviation is quite low – lower than for load-oriented order release (Wiendahl 1995), which is used as reference, but the earliness/lateness of a small number of orders is high. A number of factors may contribute to this – only the aggregate clearing function is used (no

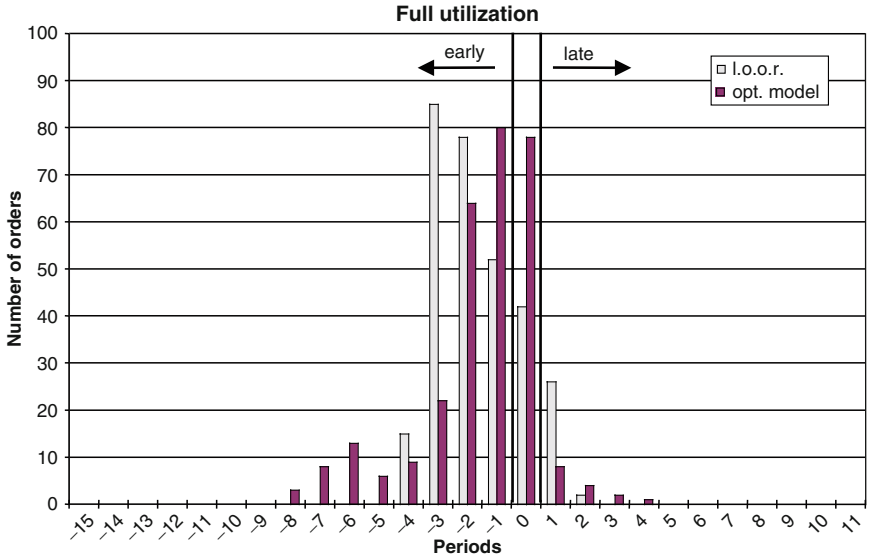

Fig. 16.13 Distribution of the due-date deviations (l.o.o.r.: load oriented order release)

partitioning; see Sect. 16.5.1), and non-bottlenecks are represented as delay functions (Missbauer 1998, p. 267 ff). However, since the clearing function only partly reflects the dynamic characteristics of the system and leads to counter-intuitive optimization results, it can be assumed that the due-date deviations in Fig. 16.13 are at least partly due to the shortcomings of the clearing function and the resulting nervousness. Future research must clarify to what extent this conclusion holds.

It is reasonable to assume that the performance of clearing function models can be improved if the history of the arrival and departure of orders at the work centers is analyzed in more detail – the clearing function aggregates the history to one dimension. Andersson et al. (1981) decompose the expected load into two parts: the expected WIP at the beginning of period t and the expected input in period t. A two-dimensional clearing function is formulated as:

$$X_t = \gamma\, W_{t-1} + \beta R_t, \tag{16.65}$$

if production does not exceed available capacity. Production, WIP and released work are measured in units of value. It can be argued that the expected WIP at the beginning of the period is actually available in period t with higher probability than the expected input during the period, which seems to be the reasoning behind this formulation. Numerical experiments based on analytical expressions for the transient $M/M/1$ queue confirm this (Missbauer, forthcoming) but the linear function in (16.65) is not derived explicitly from theory.

Conceptually, a clearing function expresses the expected WIP level that is required to obtain a certain output rate given the system variability and the production

control policies that are applied. (Anli et al. in press) present a model for order release planning that takes into account load-dependent lead times that result from the stochastic material flow, and also considers lower bounds on the finished goods inventory that are required to maintain a desired service level in the face of stochastic demand process. The lower bound on the finished goods inventory (FGI) for each SKU is a nonlinear function of the planned production volumes (production targets) of the facilities that produce and require this SKU, and of parameters representing system variability and control policies. Likewise, the expected WIP level for each unfinished SKU, facility, and period is a nonlinear function of the planned production volumes (of all SKU's) of the facility and the variables representing system variability and control policies. Both nonlinear functions can be estimated either by simulation or by queuing models. Anli et al. (in press) present their paper as a proof-of-concept study and use queuing models, namely mean value analysis. The Queuing Network Analyzer is used in Caramanis et al. (2001). Optimization is performed iteratively. In each iteration the linear constraint set of the planning model is augmented using hyperplanes tangent to the nonlinear functions. These tangents are obtained from the tentative production targets (from the previous iteration) and from the required WIP and FGI levels and their sensitivities with respect to the production targets. The authors state that this iterative refinement of the local approximations leads to convergence under mild convexity or quasi-convexity conditions.

The approach can be classified as a WIP-oriented model since it traces the flow of WIP through the facilities and does not assign flow times to the orders. The functional relationship between WIP and the production targets (volumes) can be interpreted as a sophisticated clearing function formulation that addresses the product mix problem for which we have already presented the partitioning approach (see Sect. 16.5.1) However, the model is limited to steady-state relationships and does not consider transient effects. The paper provides optimization results, but no simulation experiments. It remains to be seen whether the approach can be applied efficiently in an environment where analytical models of the manufacturing system cannot be applied.

Our discussion of the limitations of clearing function models has been limited so far to single-stage systems. Another set of complications emerges when multiple stage systems are considered. Let us assume that we wish to derive a clearing function for a work center that is part of a multistage production system – i.e., the pattern of arrivals at the workstation over time depends on production and release decisions at other work centers. Jackson (1957) in his seminal paper showed that in an open Jackson queuing network in steady state each work center can be treated as an independent $M/M/s$ queuing system, but real-world manufacturing systems often do not meet these assumptions. Therefore, in a multistage production system there are likely to be correlations between the decisions at upstream stages and the pattern of arrivals at a downstream center, which will influence the shape of the clearing function. To illustrate the point, consider a single resource that can be modeled as a $G/G/1$ queuing system in steady state. The expression (16.66) describes the average number in system $E[Ls]$ as a measure of the expected WIP for a single server where the coefficient of variation for interarrival time and service time are denoted

by c_a and c_s, respectively (see Medhi (1991) for the derivation), and ρ denotes the utilization of the server.

$$E[L_s] = \frac{c_a^2 + c_s^2}{2} \frac{\rho^2}{1-\rho} + \rho = \frac{c^2 \rho^2}{1-\rho} + \rho \tag{16.66}$$

Solving for ρ and assuming $c = c_a^2 + c_s^2 > 1$, we obtain utilization as a function of WIP as:

$$\rho = \frac{\sqrt{(E[L_s]+1)^2 + 4E[L_s](c^2-1)} - (E[L_s]+1)}{2(c^2-1)} \tag{16.67}$$

If we consider the utilization as a surrogate measure of output, Fig. 16.14 illustrates the relationship for different c values, where c combines the coefficients of variation of the arrival and service (production) processes as seen in (16.67).[6] For a fixed c value, utilization, and hence throughput, increases with WIP but at a declining rate. This is because as WIP increases, the server becomes less likely to starve. Utilization decreases as c increases, due to variability in service time and interarrival time, which causes queues to build up and throughput to slow as customers are trapped behind a customer with an unduly long service time, or the number of customers arriving in a small time interval is unexpectedly high. Note that in a multistage system, the coefficient of variation of the arrival stream c_a will be determined at least in part by the production and lot sizing decisions made at the upstream stages (e.g., giving priority to orders with low WIP at the next work center). Thus the decisions made by

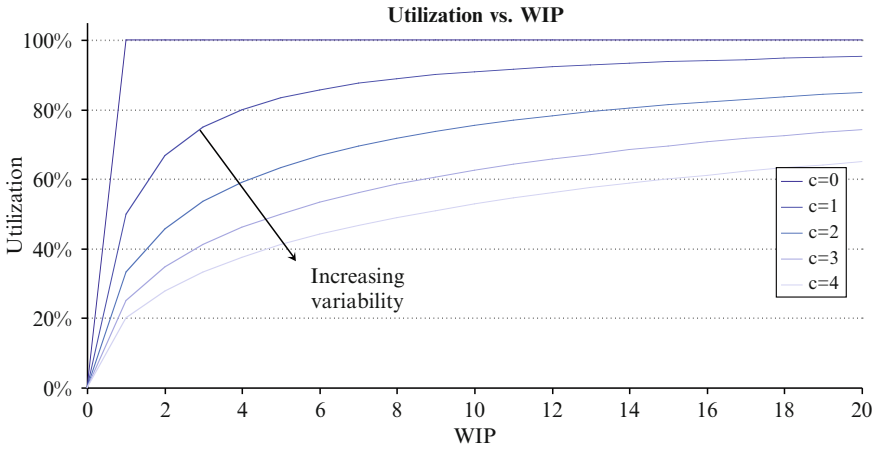

Fig. 16.14 Utilization (ρ) as a function of average WIP for different c values

[6] We are aware of the similar $GI/G/1$ approximation by Krämer and Langenbach-Belz (1976) that distinguishes between $c_a^2 \leq 1$ and $c_a^2 > 1$ (Tijms 1994), where c_a and c_s are not additive.

the model at one stage of the system affect the shape of the clearing function faced by the model at another stage. Clearing functions determined by empirical means are less susceptible to this criticism, as the correlations are captured at least in part in the data to which the functions are fitted.

16.6 Modeling Capacity with Multiple Products: Extensions to the Basic Model

In many production environments there are a number of alternative processes by which a product can be produced. These generally arise from the presence of a number of alternative machines that are capable of performing a given operation required by a product. In general, the costs of production may depend on the specific choice of equipment made for each stage. In addition, not all alternative machines for a given operation are equally efficient; a typical scenario is that there is newer equipment that can perform a given operation faster than the older equipment. Another typical scenario in high-technology industries such as semiconductor manufacturing is that the newer equipment can perform a wider range of operations than the older equipment, since the older equipment is incapable of meeting the finer tolerances that the newer equipment can handle.

In this situation, determining the optimal allocation of products to equipment over time becomes a complex decision. The dependence of costs on the particular sequence of operations followed means that a model must keep track of how much work is allocated to each possible sequence of operations, i.e., each possible path that a product can follow through the plant, as illustrated in Fig. 16.15. This figure represents a production system with four stages and a number of alternative machines, represented by the boxes. The decision variable R_{ij} denotes the amount of product i released for processing on operation sequence j. These can be

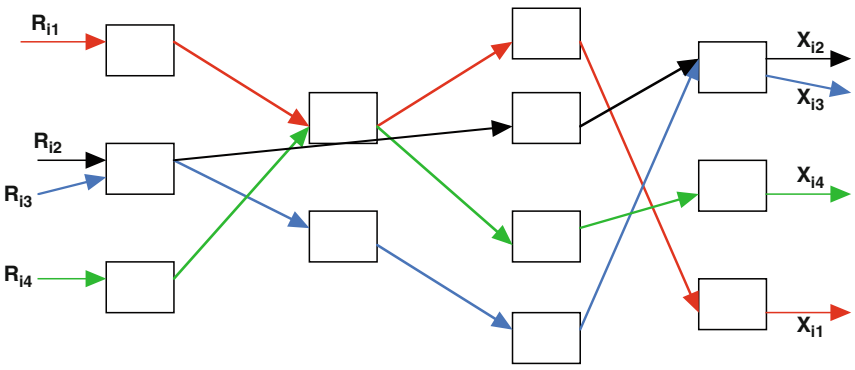

Fig. 16.15 Path-based formulation of alternative resources

considered path-based formulations, since we are explicitly specifying the amount of each product that will be launched on each possible path through the production system in each period.

These types of models, referred to as process selection models by Johnson and Montgomery (1974), have been known for quite some time, but have the obvious drawback that in a production system of any complexity, the number of possible paths through the system that a product can follow, and hence the number of decision variables in a time period, grows exponentially in the number of alternatives at each stage and the number of stages. As pointed out by Leachman and Carmon (1992), they also give us a great deal of redundant information. In order to implement the results of this model, all we need to know is the total releases of each product into the line in each period; we do not need to know the specific allocation of work to individual machines, since this is likely to change as the shop floor reacts to local circumstances and reallocates work among alternative machines. Hence a number of authors have developed models that are more compact in terms of the number of decision variables. It is not surprising that much of this work has been motivated by applications in the semiconductor industry, where a product may require several hundred unit processes, all of which may have a significant number of alternative machines.

Much of the work in this area takes as its starting point the LP models with fixed time lags treated by Hackman and Leachman (1989). Leachman and Carmon (1992) present a series of models that address this issue. We shall focus on this paper in some detail, and then outline the extensions proposed by other authors. We define the following notation:
Parameters:

a_{ijk} = time to process one unit of product i at step j on machine type k.
C_{kt} = capacity in time unit of machine type k in period t.
D_{it} = maximum cumulative demand for product i in period t, made up of the forecast and confirmed orders.
d_{it} = minimum cumulative demand for product i by time t (firm orders).
p_{it} = estimated net discounted cash flow from selling one unit of product i in period t.
h_{it} = estimated unit holding cost for product i in period t.
L_i = average flow time for product i from beginning to end of its entire process, i.e., the cycle time for the entire process.
L_{ij} = average lead time for product i from start of production process until start of step j.

The decision variables, which are common to all the three formulations presented, are as follows:

R_{it} = number of units of product i to be released in period t.
I_{it} = units of product i in inventory in period t.
B_{it} = shortfall of cumulative production vs. cumulative max demand for product i in period t.

The first model, referred to by the authors as the Step-Separated Formulation, introduces additional decision variables W_{ijkt} that denote the amount of workload of each product type i in each process step j that is assigned to machine type k in period t. The model can now be stated as follows:

$$\max \sum_{i=1}^{n} \sum_{t=1}^{T} [p_{it} R_{i,t-L_i} - h_{it} I_{it}], \tag{16.68}$$

subject to

$$R_{i,t-L_{ij}} = \sum_{k \in P(i,j)} \frac{W_{ijkt}}{a_{ijk}}, \quad \text{for all } i, j, t, \tag{16.69}$$

$$\sum_{\{(i,j)|k \in P(i,j)\}} W_{ijkt} \leq C_{kt}, \quad \text{for all } k, t, \tag{16.70}$$

$$\sum_{\tau=1}^{t} R_{i,\tau-L_i} - I_{it} + B_{it} = D_{it}, \quad \text{for all } i; \text{for all } t \leq T-1, \tag{16.71}$$

$$\sum_{\tau=1}^{T} R_{i,\tau-L_i} + B_{iT} = D_{iT}, \quad \text{for all } i, \tag{16.72}$$

$$B_{it} \leq D_{it} - d_{it}, \quad \text{for all } i, t, \tag{16.73}$$

$$R_{it}, B_{it}, I_{it}, W_{ijkt} \geq 0. \tag{16.74}$$

The critical constraints for the purposes of modeling capacity and workload are the first two. The first set of constraints converts the releases in period $t - L_{ij}$ into the current workload on the resources k in period t. The second set of constraints then ensures that no resource is loaded in excess of its capacity. Note that accumulation of WIP within the production process is not modeled at all; only finished goods inventory is represented in the decision variables. The objective function is to maximize the difference between the discounted revenue from sales and inventory holding costs. Note that a product is credited as producing revenue as soon as it is produced, even if it may not be sold immediately. Also note that in the last two constraint sets, output aimed at meeting forecasts may be delivered late, but firm orders must be met on time. This effectively creates two hierarchical demand classes, with the demand from firm orders having absolute priority over forecast demand in terms of allocating limited output to meet demand. Production costs are not considered, as they are independent of production routing and the revenue is assumed to be much greater than unit production cost, as is the case in much of the semiconductor industry. Hence the emphasis of the model is on allocating production capacity to meet demand, and thus maximize revenue.

This formulation is inefficient in that it gives a detailed allocation of workload to individual stations (the W_{ijkt} variables), even though all we really need is the total releases R_{it} of each product in each period. In order to arrive at a more efficient

formulation, the authors make the additional assumption of uniform processing times across machines in a work center, i.e., $a_{ijk} = a'_{ij}S_k$, where a'_{ij} denotes the processing time of product i at step j on the standard machine whose speed is used as a baseline for the others. The available capacity of each set of alternative machines can then be rescaled in a similar manner, $C'_{kt} = C_{kt}/s_k$. Now since all workloads are expressed in terms of the baseline machine, we only need to track the total workload at step j that is assigned to machine type k. In order to do this, we define unique sets S_m of alternative machine types, and additional variables Z^m_{kt} that denote the workload on machine set S_m assigned to machine type k in period t. Defining the set $P(i, j)$ as the set of machines capable of processing step j of product i, this yields the following formulation:

$$\max \sum_{i=l}^{n} \sum_{i=l}^{n} [\rho_{it} R_{i\,t-l_i} - h_{it} I_{it}], \tag{16.75}$$

subject to

$$\sum_{\{(i,j)|k\in P(i,j)\}} a'_{ij} R_{i,t-L_{ij}} = \sum_{k\in S_m} Z^m_{kt}, \quad \text{for all } m, t, \tag{16.76}$$

$$\sum_{m=1}^{M} Z^m_{kt} \le C'_{k\,t}, \quad \text{for all } k, t \tag{16.77}$$

$$\sum_{\tau=1}^{t} R_{i,t-L_i} - I_{it} + B_{it} = D_{it}, \quad \text{for all } i; \text{for all } t \le T-1, \tag{16.78}$$

$$\sum_{\tau=1}^{T} R_{i,t-L_i} + B_{iT} = D_{iT}, \quad \text{for all } i, \tag{16.79}$$

$$B_{it} \le D_{it} - d_{it}, \quad \text{for all } i, t, \tag{16.80}$$

$$R_{it}, B_{it}, I_{it}, Z^m_{kt} \ge 0. \tag{16.81}$$

Using these assumptions, the authors construct a formulation where all allocation variables are eliminated, and capacity constraints are written for sets of alternative machines whose capacity is likely to be binding on the optimal solution. The structure of these sets depends on the problem data, and hence this formulation is not always the most compact in terms of the number of variables and constraints, but in many industrial situations yields substantially smaller formulations. The sets S of alternative machines for whom capacity constraints will be written are determined based on cut sets in the bipartite graph representing operation-machine requirements. The authors provide a procedure to identify the dominant cut sets whose time complexity is linear in the number of operations ij and the cardinality of the alternative machine sets, but exponential in the number of machines that occur in the connected component of the bipartite graph. Thus, if there are a very large number of machines that can process a large number of steps as alternatives,

the time complexity of generating this formulation may be quite high, but such instances seldom arise in industrial data sets. Once the dominant cut sets S have been obtained, we can write capacity constraints for each set and each period of the form

$$\sum_{ij \in S} a_{ij} R_{i,t-L_{ij}} \leq \sum_{k \in S} C_{kt}. \tag{16.82}$$

The complete formulation is now as follows:

$$\max \sum_{i=1}^{n} \sum_{t=1}^{T} [p_{it} R_{i,t-L_i} - h_{it} I_{it}], \tag{16.83}$$

subject to

$$\sum_{ij \in S} a_{ij} R_{i,t-L_j} \leq \sum_{k \in S} C_{kt}, \quad \text{for all generated sets } S \text{ and periods } t, \tag{16.84}$$

$$\sum_{\tau=1}^{t} R_{i,t-L_i} - I_{it} + B_{it} = D_{it}, \quad \text{for all } i; \text{for all } t \leq T-1, \tag{16.85}$$

$$\sum_{\tau=1}^{t} R_{i,\tau-L_i} + B_{iT} = D_{iT}, \quad \text{for all } i, \tag{16.86}$$

$$\sum_{\tau=1}^{T} R_{i,\tau-L_i} + B_{iT} = D_{iT}, \quad \text{for all } i, \tag{16.87}$$

$$R_{it}, B_{it}, I_{it} \geq 0. \tag{16.88}$$

The authors analyze the number of decision variables and constraints in their different formulations and show that when alternative machine sets have a nested property, which occurs frequently in semiconductor manufacturing, this Direct Product Mix formulation provides a very compact model compared to alternative formulations. The nested property arises when a work center has machines of several technological generations, where each newer generation can perform all operations the previous generations could, as well as some additional new ones.

A drawback of the Direct Product Mix formulation developed above is its reliance on the assumption of uniform processing times across alternative machines as described above. Bermon and Hood (1999), in their study of production planning for IBM's semiconductor manufacturing operations, noted that this assumption was violated in their environment. Hung and Cheng (2002) extend the work of Leachman and Carmon (1992) by developing a formulation that does not require the uniformity assumption. In order to do this, they define a new set of partitioning variables that allocate the capacity of machines shared between machine sets to which they belong. Their computational experiments show that when the uniformity assumption on processing times holds, the Direct Product

Mix approach of Leachman and Carmon (1992) is preferable. However, when the uniformity assumption is violated, the Partition formulation developed by the authors remains valid.

Bermon and Hood (1999) present a slightly different model aimed at capacity planning that also addresses the problem of determining capacity in situations with alternative machine sets. Hung and Wang (1997) apply an approach similar to that of Leachman and Carmon (1992) to situations where alternative products can be used to meet a given demand, as by downgrading in electronics manufacturing.

16.7 Lot Sizing Models

As we have seen, the clearing function reflects the variability of the arrival and departure process. If a work center produces multiple products in lots, lot sizing determines the operation times of the orders and thus strongly influences the variability of the operation times. It also determines the total setup time during a period and thus influences the maximum output of the work center. Hence the lot sizes influence the average flow times and the WIP level at the work centers.

Considering the impact of lot sizes on average flow time and WIP means to anticipate consequences that become visible at the scheduling level. One way to achieve this is simultaneous lot sizing and scheduling. This economic lot scheduling problem (ELSP) has been studied extensively (for reviews, see Elmaghraby 1978; Graves 1981; Drexl and Kimms 1997; Pinedo and Chao 2005). Drum-Buffer-Rope-OPT, which has been described extensively in the 1980s, also performs simultaneous lot sizing and scheduling for the bottleneck work centers, distinguishing between *transfer* and (often larger) *process batches.* See Zäpfel and Missbauer (1993b) for related literature on OPT.

In accordance with the hierarchical structure of the manufacturing planning and control system that we assume throughout the chapter, detailed scheduling is performed locally within the production units. Thus we do not consider simultaneous lot sizing and scheduling or cyclic production. We assume that lot sizing passes the lots (production orders) to the order release function where they are released to the production units that perform sequencing. In this context, stochastic models of the manufacturing system are appropriate to anticipate the consequences of lot sizes on flow times and WIP.

Assuming a stationary state of the system, we can examine the relationship between the lot sizes and the long-term clearing function of a work center by means of an $M/G/1$ model producing products with identical data. We define m as the total demand rate (sum of the identical demand rates of the products), a the processing time per unit, r the setup time per lot, σ and v the standard deviation and coefficient of variation of the service times of the orders, x the lot size. a, r, and x are identical for all products. We assume that a setup is necessary for each lot. Considering setup time savings obtained by sequence optimization is described in Kekre (1984), Kekre (1987) and Missbauer (1997).

For simplicity we assume a given value for the coefficient of variation of the service times v that is independent of the common lot size x. This is sufficient for showing the structural properties. Missbauer, (2002b) and Karmarkar et al. (1985a, 1985b) discuss the relationship between lot sizes and the coefficients of variation of service times in the multi-product case. We also assume a Poisson arrival process, which is a rather strong assumption in the single-product case (see Kistner (1999) for a critique), but in the multi-product case with identical products this assumption is well justified (Missbauer 1999). Based on these assumptions we compute the arrival rate

$$\lambda = \frac{m}{x} \tag{16.89}$$

and the mean service rate

$$\mu = \frac{1}{r + ax}. \tag{16.90}$$

The mean waiting time $E[W_q]$ by the Pollaczek–Khinchine formula is then

$$E\left[W_q\right] = \frac{\rho^2 + \lambda^2\sigma^2}{2\lambda(1-\rho)} \tag{16.91}$$

with $\rho = \lambda/\mu = m(a + r/x)$, and the mean flow time

$$E[W_s] = E[W_q] + 1/\mu. \tag{16.92}$$

Substituting (16.89) for λ, (16.90) for μ and $\sigma = v(r + ax)$ into (16.91), we get the average waiting time as

$$E[W_q] = \frac{m(v^2+1)(ax+r)^2}{2[x(1-am) - mr]}. \tag{16.93}$$

Because of the assumption of Poisson input and the PASTA (Poisson arrival see time averages) property (Buzacott and Shanthikumar 1993, p. 54; Tijms 1994, p. 73 ff.) the average waiting time of the customers $E[W_q]$ must be identical to the average WIP at the server, measured in hours of work, given by the average remaining work $E[L_w]$). So we can write

$$E[L_w] = E[W_q]. \tag{16.94}$$

The average output in hours of work as a function of the average WIP $E(L_w)$ can be calculated from (16.93) and (16.94) as follows:

$$\text{Output} = m \cdot a = \frac{2axE[L_w]}{(ax+r)\left[2E[L_w] + (v^2+1)(ax+r)\right]} \tag{16.95}$$

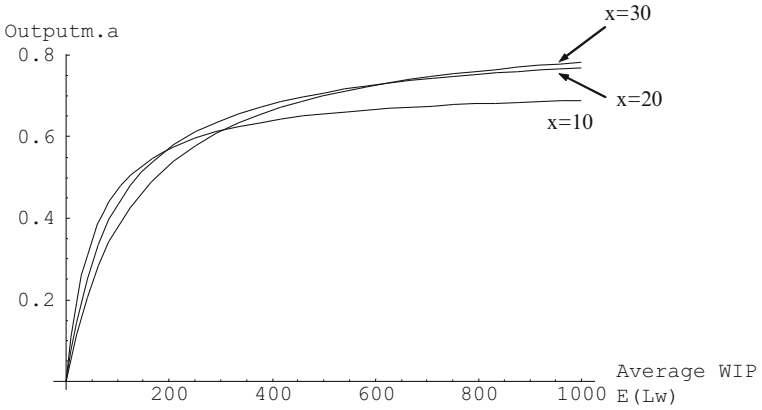

Fig. 16.16 Output as a function of WIP (16.95) for different values for the lot size $x (a = 5, r = 15)$

From (16.95) we see that a higher variability of the service times decreases the output for a given average WIP. The clearing function for different lot sizes is shown in Fig. 16.16.

This adds an important aspect to the lot sizing problem. Lot sizing should not only consider setup and inventory holding costs but also its impact on the clearing function, which can be regarded as capacitated lot sizing considering congestion effects. There are two ways to accomplish this

- Determination of standard lot sizes or lot sizing rules that take into account the impact of lot sizes on the clearing function (and hence on WIP and flow times that result from order release). The clearing function that results from the lot sizes is used for order release. In the extreme case lot sizes can be determined such that average flow time or WIP is minimized for the required output. (Note that usually this is in conflict with the traditional goal of minimizing the sum of setup and holding costs). Starting with (Karmarkar et al. 1985a, b; Karmarkar 1987) the majority of the literature on this topic is based on this idea (Missbauer 2002b).
- Defining a multi-dimensional clearing function with output as a function of WIP and lot sizes as independent variables:

$$X_{it} \leq f_i \left(W_{it}, x_{1t}, x_{2t}, \ldots, x_{Nt} \right). \tag{16.96}$$

  In this case the order release model optimizes the time-varying lot sizes and the amount of released work for all periods simultaneously. This recent research direction is presented in (Hwang and Uzsoy 2005). Note that the steady-state assumption implies that the changes in the characteristics of the arriving orders affect the clearing function within the same period.

Since lot sizes influence lead times, they also influence the lead time demand distribution, which in turn influences optimal safety stocks and reorder points. For

this aspects, see Lambrecht et al. (1996) and Vaughan (2006). Little is known about multistage lot sizing considering congestion effects. For a model exploring this issue, see Missbauer (1999, 2002b).

16.8 Models Incorporating Uncertainty

Until this point we have treated all parameters of our various optimization formulations as deterministic, although in some cases they represent the expectation of a performance measure derived from an underlying stochastic system. However, in any industrial application all parameters of an optimization model, such as the cost estimates used in the objective function, the technological coefficients defining resource consumption by products, and especially the forecasts of future demand, are subject to significant uncertainties. The explicit treatment of uncertainty in optimization formulations of production planning problems is very much in its infancy, so we will focus on illustrating the issues in one particular area – that of uncertainty in demand forecasts, which requires the system to hold a certain amount of safety stock to maintain a given level of customer service.

It is well known that the amount of safety stock that needs to be held in a particular location to maintain a specified level of customer service is related to the distribution of the demand forecast over the replenishment lead time. To illustrate this relationship, consider a simple newsvendor model where the lead time is a random variable with mean μ_L and variance σ_L^2, and the demand rate has a mean of μ and a variance of σ^2. It is well known (e.g., Eppen and Martin 1988) that the optimal order level can be approximated by

$$\mu_L \mu + z_\alpha \sqrt{\mu_L \sigma^2 + \sigma_L^2 \mu^2}, \tag{16.97}$$

assuming that the lead time is normally distributed. Note that in the term under the square root, which denotes the standard deviation of the demand over the lead time, both the mean and the variability of the lead time interact with the mean and variability of the demand to influence the amount of safety stock required. However, the lead times, in turn, are determined by the utilization levels of the resources. While several approaches described in the previous section address this issue in terms of planning models, there have been few efforts to address this directly in the planning literature, although several stochastic models linking queues with inventory models have been proposed (e.g., Zipkin 1986; Liu et al. 2004).

In most current approaches in industry, many of which have their origin in the MRP literature, the amount of safety stock to be held in a particular location in a particular planning period is calculated outside of the planning model using actual or estimated parameters of the demand or demand forecast distribution. The planning model is then constrained to maintain this quantity of safety stock. Examples of such models are described at length by Wijngaard and Wortmann (1985) in the context of MRP systems, where the amount of safety stock to be maintained results

in the release of additional orders to the system. Lambrecht et al. (1984a) show that the problem of determining the amount of safety stock in multistage production systems can be formulated as a Markov decision process or a dynamic program, but the size of the state space renders these approaches impractical from a computational point of view. They propose a number of heuristics, and conduct extensive computational experiments that examine the amount of safety lead time and safety stock required. Lambrecht et al. (1984b) perform computational experiments with the Markov decision process and examine the form of the optimal policies. Yano and Carlson (1988) propose a heuristic for setting safety stock in an assembly-type production system and examine its performance using simulation.

Graves (1986, 1988) develops an intriguing model that expressly links safety stock levels to the ability of the production system to react to changes, using a model that separates WIP and FGI and uses the proportional clearing function in Fig. 16.5 to model the production behavior of the facility based on planned lead times. However, his model assumes a stationary demand distribution, and the proportional clearing function may produce capacity-infeasible solutions at high WIP levels. He points out that distinguishing between WIP and FGI is significant since in many production systems WIP can serve some of the function of safety stock. This indicates the desirability of making safety stock decisions endogeneous in a production planning model that combines a realistic aggregate model of the behavior of congestion-prone capacitated production systems (such as that given by the clearing function formulations) with the explicit modeling of WIP and FGI as separate entities. We conjecture that such a model might well be capable of maintaining a given service level with significantly less safety stock in the form of finished goods between stages, since it would be able to recognize that WIP in the line would become available to meet demands, reducing the need for stocks of finished goods.

Leachman (1993) presents a large-scale linear programming framework for production planning in the semiconductor industry where safety stocks are addressed through the use of demand classes. Production required to replace safety stocks is modeled as a class of demand that has lower priority than firm customer orders, and capacity is allocated to these orders only if this will not compromise the ability of the system to meet firm customer orders. He suggests simulation of the production plan to obtain estimates of the variability of the resulting lead times this plan will impose on the system. These estimates of variability can then be used to determine safety stock levels to protect against variability in lead times. Hung and Chang (1999) elaborate further on this approach and give computational experiments examining its performance.

In the domain of production planning models, the chance-constrained (CC) formulations of Charnes and Cooper (1963) can be used to obtain deterministic equivalents to a number of production planning problems involving random variables. In the simplest version of this approach, consider demand to be the only source of uncertainty, and assume that demand in period t has a cumulative distribution function F_t. Note that I_t is a random variable due to the demand being random. Let us define U_t to be a target inventory level such that in each period t we produce $X_t = U_t - I_{t-1}$ units if $U_t > I_{t-1}$ and otherwise do not produce in that period.

Then we can write chance constraints of the form $P\{I_t \geq 0\} \geq 1-\alpha$, where α denotes the acceptable probability of stockout in a period. The optimization problem is to determine the X_t. Defining $g_t = U_t - U_{t-1}$, implying $X_t = g_t + D_{t-1}$, yields

$$I_t = I_0 + \sum_{\tau=1}^{t} (X_\tau - D_\tau) = I_0 + \sum_{\tau=1}^{t} g_\tau - D_t. \tag{16.98}$$

This allows us to write the chance constraint as

$$P\left\{ I_0 + \sum_{\tau=1}^{t} g_t \geq D_t \right\} \geq 1-\alpha \tag{16.99}$$

implying

$$I_t = I_0 + \sum_{\tau=1}^{t} g_\tau \geq F^{-1}(1-\alpha), \tag{16.100}$$

where the right-hand side is now a constant. Most existing CC production planning models are uncapacitated, and do not model WIP in any form. The chance constraints are thus developed to ensure that the finished goods inventory at the end of each period is positive with a given probability, corresponding to the service level desired. The most common approach to developing an objective function is to assume that the probabilities of constraint violation are sufficiently small that backorder costs can be neglected, as suggested by Bookbinder and Tan (1988) and Johnson and Montgomery (1974). Similar CC formulations of chance-constrained problems are given by Bookbinder and H'ng (1986), Gupta and Sengupta (1977), Sengupta (1972), Sengupta and Portillo-Campbell (1973), and Rakes et al. (1984), among others.

This approach does not appear to have been used much in recent years; stochastic programming (see, e.g., Birge and Louveaux 1997) has been preferred, for strong reasons related to the difficulties of the chance-constrained approach in modeling recourse actions (Blau 1974; Hogan et al. 1981; Charnes and Cooper 1983). The stochastic programming formulation is mathematically more complete in terms of its ability to model multiple stage decision problems with recourse actions possible at each stage. However, the large number of decision stages, corresponding to the time periods in planning problems encountered in industry, renders the use of stochastic programming computationally challenging, as suggested by Peters et al. (1977). The CC formulation has a number of difficulties – the desired probabilities of constraint violation need to be specified *a priori*, and the degree to which the constraints are violated is not accounted for in the objective function. From a practical perspective, the models are infeasible if it is not possible to satisfy all chance constraints with the desired probabilities, without providing the user any means of trading off service levels between products. In order to obtain tractable constraint sets, distributional assumptions must be made about the random variables on the

right hand sides of the constraints. Extensive discussions of these issues can be found in, for example, Prekopa (1993) and Lejeune and Prekopa (2005).

However, the CC formulations also offer a number of advantages for practical implementation relative to stochastic programming, The first of these is that with varying degrees of approximation, depending on the degree to which the distributional assumptions on the random variables are violated, these models can be implemented using extensions of the LP formulations with which both practitioners and researchers are familiar. While pre-specifying the probabilities of constraint violation may be problematic in many application domains, in the context of production planning that we consider, the probability of constraint violation has a natural interpretation as the probability of a stockout. The need to pre-specify stockout probabilities may actually be an advantage in practice, as it forces users to think in terms of service levels, perhaps based on aggregating products into product families or customers into priority classes.

A specific kind of demand uncertainty can emerge in production planning models that decide on *aggregate* sales, production and inventory, usually over the seasonal cycle (sales and operations planning; see Vollmann, Berry et al. 2005, p. 60 ff.). Typically the products are aggregated into groups or families of products with similar demand pattern and resource requirements, and the decision variables are defined at this aggregate level. Since stockouts are defined at the level of individual products, nonnegativity of the aggregate inventory is not a sufficient condition for a feasible production plan. If the aggregate demand is considered as deterministic and the demand of individual products is uncertain with lower and upper bounds, the task is to find a *robust* aggregate production plan that allows a feasible solution at the level of individual products (disaggregation to obtain a feasible master production schedule). For this topic, see Lasserre and Mercé (1990) and Gfrerer and Zäpfel (1995).

16.9 Conclusions and Future Directions

The problems discussed in this chapter, of managing the release of work into a production system and allocating resource capacity among different products, constitutes one of the earliest applications of operations research to industrial planning and control problems, with literature dating back more than five decades. When viewed as part of a production planning and control hierarchy, the workload control approaches developed over the years focus on maintaining predictable lead time and throughput behavior in a stable demand environment, where the system can be expected to produce at a relatively constant rate. Traditional order release mechanisms, complemented by suitable methods for order acceptance and due date setting, are also recommended for make-to-order production where demand forecasts are difficult to obtain. This implies that the objective of these methods is to maintain a desirable pattern of aggregate material flow through the facility, which these techniques try to accomplish mainly by heuristic means with relatively little unifying theoretical support. The lack of a strong theoretical understanding of this

area is evidenced by the fact that most of the existing knowledge from this type of research is in the form of results from simulation studies, which are sometimes contradictory, and often hard to generalize beyond the specific production system topology in which they were tested.

Focusing on optimization of aggregate material flows through the production system motivates the discussion of the second, higher level of the planning hierarchy, where a more aggregate plan for work release and capacity allocation take place. These models generally take a more aggregate perspective, with time being divided into discrete planning periods and material flows being viewed as a continuous medium as opposed to discrete jobs that must be handled as an integral unit. We have focused in particular on the rich literature on mathematical programming models of these problems, almost all of which can trace their ancestry to the work of Holt, Modigliani and their collaborators in the 1950s (Holt et al. 1955, 1956, 1960; Modigliani and Hohn 1955). It is interesting to note that until very recently, there has been a hiatus in research on these models; between the late 1970s and the late 1990s there are relatively few papers on formulation and modeling aspects of these problems, with the work of Leachman and his coworkers (Hackman and Leachman 1989; Leachman and Carmon 1992; Hung and Leachman 1996; Dessouky and Leachman 1997) being a significant exception. It is also interesting to note that the 1974 book by Johnson and Montgomery is still one of the best available references for most of the classical work in this area. One is left with the feeling that for many years this area was perceived as a "solved" problem with no further interesting research issues.

We hope that the discussion in this chapter will stimulate wider interest in both industry and academia in this area. While widely taught in academia and used in industry, the classical linear programming models have a number of limitations arising from their very aggregate, static approach to modeling production capacity. It is heartening that in recent years a growing number of researchers have begun to explore these problems anew (Pahl et al. 2005). The new approaches differ substantially from the classical approaches in their efforts to achieve solutions that are consistent with the queuing behavior of production systems, which is well studied (Buzacott and Shanthikumar 1993; Hopp and Spearman 2001), and thus tend to have nonlinear structure which, in many cases, can be addressed effectively in computational procedures.

A number of important research directions have been outlined in the chapter, but are worth summarizing again. The new nonlinear approaches, among which the clearing function approach appears to be the most studied, show considerable promise but need to be better understood both empirically and theoretically. The issues of how to derive clearing functions analytically in multistage systems when decisions at one stage affect the variability of arrivals, and hence the shape of the clearing function, at downstream stages needs to be examined. There also needs to be a better understanding of the implications of using steady-state queuing results to develop clearing functions for use in a dynamic, nonstationary environment, where the purpose of the planning process is to change release rates over time. A closely related and not well-understood issue is that of how changes in decision variable

values at the boundaries between planning periods affect the implementability and execution of the plans obtained. To what extent insights on the transient behavior of queuing systems should be integrated into the models is not known today. In terms of empirical estimation of clearing functions, we have experimental results demonstrating that a simple least-squares fit to empirical data may result in very poor planning models, but there is no theoretically justified alternative approach available as yet.

The main alternative to clearing function models are lead time-oriented models. Fixed lead times do not recognize the load-dependence of the lead times in the case of time-varying capacity load. A fixed relationship between lead time distribution and capacity load in a period can lead to substantial modeling errors since it does not capture the dynamic characteristics of lead times. The iterative approaches of Hung and Leachman (1996) and Riaño et al. (2006) are very interesting, but their computational performance, especially their convergence characteristics, have not been tested extensively.

Similar concerns hold for most of the other approaches that have been suggested as alternatives. While both stochastic programming and chance constraint formulations have been proposed for addressing the issue of uncertainty inherent in most industrial applications, effective computational procedures are not available, and the implications of the formulations are not well understood. Most research on lot-sizing has focused on developing effective solution procedures for the resulting fixed-charge integer programming models, but the majority of these models use the same model of capacity as the classical linear programming models, and there is clearly much work to be done here.

Finally, from the point of view of industrial applications, it is notable that many of the proposed new approaches are significantly more complex in both their data and their computational requirements, and especially load-dependent lead times may complicate coordination between manufacturing departments. It is by no means obvious that the proposed new models are always superior to the classical models in all industrial environments. This requires developing a better, theory-based understanding of the conditions under which the additional complexity of the new models is justified over the well-understood classical models that have been the mainstay of industrial practice for several decades.

Acknowledgments The research of Reha Uzsoy was supported by the National Science Foundation under Grant DMI-0556136, by the Intel Research Council, by a software grant from Dash Optimization and an equipment grant from Intel Corporation.

References

Agnew C (1976) Dynamic modeling and control of some congestion prone systems. Oper Res 24(3):400–419

Andersson H, Axsater S et al. (1981) Hierarchical material requirements planning. Int J Prod Res 19(1):45–57

Anli OM, Caramanis M et al. (2007). Tractable supply chain production planning modeling nonlinear lead time and quality of service constraints. J Manuf Syst 26(2):116–134
Anthony RN (1966) Planning and control systems: a framework for analysis. Harvard University Press, Cambridge
Asmundsson JM, Rardin RL et al. (2006) Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. IEEE Trans Semicond Manuf 19:95–111
Asmundsson JM, Rardin RL et al. (2009) Production planning models with resources subject to congestion. Naval Res Log 56:142–157
Baker KR (1993) Requirements planning. In: Graves SC, Rinnooy Kan AHG, Zipkin PH. Logistics of production and inventory. Handbooks in operations research and management science, vol 3. Elsevier Science, Amsterdam, pp 571–627
Bergamaschi D, Cigolini R et al. (1997) Order review and release strategies in a job shop environment: a review and a classification. Int J Prod Res 35:399–420
Bermon S, Hood SJ (1999) Capacity optimization planning system (CAPS). Interfaces 29(5):31–50
Bertrand JWM, Wortmann JC (1981) Production control and information systems for component-manufacturing shops. Elsevier, Amsterdam
Bertrand JWM, Wortmann JC et al. (1990) Production control: a structural and design oriented approach. Elsevier, Amsterdam
Bertsimas D, Gamarnik D et al. (2003) From fluid relaxations to practical algorithms for high-multiplicity job shop scheduling: the holding cost objective. Oper Res 51(5):798–813
Bertsimas D, Sethuraman J (2002) From fluid relaxations to practical algorithms for job shop scheduling: the makespan objective. Math Program Series A 92:61–102
Birge JR, Louveaux F (1997) Introduction to stochastic programming. Springer, New York
Bitran GR, Haas EA et al. (1981) Hierarchical production planning: a single stage system. Oper Res 29(4):717–743
Bitran GR, Haas EA et al. (1982) Hierarchical production planning: a two-stage system. Oper Res 30(2):232–251
Bitran GR, Tirupati D (1993) Hierarchical production planning. Graves SC, Rinnooy Kan AHG, Zipkin PH Logistics of production and inventory. Handbooks in operations research and management science, vol. 4. Elsevier Science, Amsterdam, pp 523–568
Blau RA (1974) Stochastic programming and decision analysis: an apparent dilemma. Manage Sci 21(3):271–276
Bookbinder JH, H'ng BT (1986) Rolling horizon production planning for probabilistic time-varying demands. Int J Prod Res 24(6):1439–1458
Bookbinder JH, Tan JY (1988) Strategies for the probabilistic lot sizing problem with service level constraints. Manage Sci 34(9):1096–1108
Bowman EB (1956) Production scheduling by the transportation method of linear programming. Oper Res 4(1):100–103
Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice-Hall, Englewood Cliffs
Byrne MD, Bakir MA (1999) Production planning using a hybrid simulation-analytical approach. Int J Prod Econ 59:305–311
Byrne MD, Hossain MM (2005) Production planning: an improved hybrid approach. Int J Prod Econ 93–94:225–229
Caramanis M, Pan H et al. (2001) A closed-loop approach to efficient and stable supply chain coordination in complex stochastic manufacturing. American Control Conference, Arlington, VA, 1381–1388
Carey M (1987) Optimal time-varying flows on congested networks. Oper Res 35(1):58–69
Carey M, Subrahmanian E (2000) An approach to modelling time-varying flows on congested networks. Transp Res B 34:157–183
Cassidy M (2003) Traffic flow and capacity. In: Hall RW (ed) Handbook of transportation science. Kluwer Academic, Dordrecht, pp 155–191
Charnes A, Cooper WW (1963) Deterministic equivalents for optimizing and satisficing under chance constraints. Oper Res 11:18–39

Charnes A, Cooper WW (1983) Response to "decision problems under risk and chance constrained programming: dilemmas in the transition". Manage Sci 29(6):750–753
Charnes A, Cooper WW et al. (1955) A model for optimizing production by reference to cost surrogates. Econometrica 23(3):307–323
Chen HB, Mandelbaum A (1991) Hierarchical modelling of stochastic networks part I: fluid models. In: Yao DD (ed) Stochastic modeling and analysis of manufacturing systems. Springer, New York
Cohen JW (1969) The Single server queue. North-Holland, Amsterdam
Cohen O (1988) The drum-buffer-rope (DBR) approach to logistics. In: Rolstadas A (ed) Computer-aided production management. Springer, New York
Davidson R, MacKinnon JG (1993) Estimation and inference in econometrics. Oxford University Press, New York
de Kok AG, Fransoo JC (2003) Planning supply chain operations: definition and comparison of planning concepts. In: de Kok AG, Graves SC (eds) OR Handbook on supply chain management. Elsevier, Amsterdam, pp 597–675
Dessouky MM, Leachman RC (1997) Dynamic models of production with multiple operations and general processing times. J Oper Res Soc 48(6):647–654
Drexl A, Kimms A (1997) Lot sizing and scheduling – survey and extensions. Eur J Oper Res 99:221–235
Elmaghraby SE (1978) The economic lot scheduling problem (ELSP): review and extensions. Manage Sci 24:587–598
Eppen G, Martin RK (1988) Determining safety stock in the presence of stochastic lead times. Manage Sci 34:1380–1390
Fine CH, Graves SC (1989) A tactical planning model for manufacturing subcomponents of mainframe computers. J Manuf Oper Manage 2:4–34
Forrester JW (1962) Industrial dynamics. MIT Press, Cambridge
Fredendall LD, Ojha D, Patterson W (2010) Concerning the theory of workload control. Eur J Oper Res 201:99–111
Gfrerer H, Zäpfel G (1995) Hierarchical model for production planning in the case of uncertain demand. Eur J Oper Res 86:142–161
Graves SC (1981) A review of production scheduling. Oper Res 29(4):646–675
Graves SC (1986) A tactical planning model for a job shop. Oper Res 34:552–533
Graves SC (1988) Safety stocks in manufacturing systems. J Manuf Oper Manage 1:67–101
Gunther HO, Van Beek P (2003) Advanced planning and scheduling solutions in process industry. Springer, Heidelberg
Gupta M (2005) Constraints management – recent advances and practices. Int J Prod Res 41(4):647–659
Gupta SK, Sengupta JK (1977) Decision rules in production planning under chance-constrained sales. Decision Sci 8:521–533
Hackman S (2008) Production economics. Springer, Berlin
Hackman ST, Leachman RC (1989) A general framework for modeling production. Manage Sci 35:478–495
Hanssmann F, Hess SW (1960) A linear programming approach to production and employment scheduling. Manage Technol 1(1):46–51
Harris FW (1915) Operations and cost. Factory management series. Shaw, Chicago
Hax AC, Candea D (1984) Production and inventory management. Prentice-Hall, Englewood Cliffs
Haxholdt C, Larsen ER et al. (2003) Mode locking and chaos in a deterministic queueing model with feedback. Manage Sci 49(6):816–830
Hendry LC, Kingsman BG (1991) A decision support system for job release in make to order companies. Int J Oper Prod Manage 11:6–16
Hogan AJ, Morris JG et al. (1981) Decision problems under risk and chance constrained programming: dilemmas in the transition. Manage Sci 27(6):698–716
Holt CC, Modigliani F et al. (1955) A linear decision rule for production and employment scheduling. Manage Sci 2(1):1–30

Holt CC, Modigliani F et al. (1956) Derivation of a linear rule for production and employment. Manage Sci 2(2):159–177
Holt CC, Modigliani F et al. (1960) Planning production, inventories and work force. Prentice Hall, Englewood Cliffs
Hopp WJ, Spearman ML (2001) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston
Hung YF, Chang CB (1999) Determining safety stocks for production planning in uncertain manufacturing. Int J Prod Econ 58:199–208
Hung YF, Cheng GJ (2002) Hybrid capacity modelling for alternative machine types in linear programming production planning. IIE Trans 34:157–165
Hung YF, Hou MC (2001) A production planning approach based on iterations of linear programming optimization and flow time prediction. J Chinese Inst Ind Engrs 18(3):55–67
Hung YF, Leachman RC (1996) A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations. IEEE Trans Semicond Manufac 9(2):257–269
Hung YF, Wang QZ (1997) A new formulation technique for alternative material planning – an approach for semiconductor bin allocation. Comput Ind Eng 32(2):281–297
Hwang S, Uzsoy R (2005) A single stage multi-product dynamic lot sizing model with work in process and congestion. Research report, Laboratory for Extended Enterprises at Purdue, School of Industrial Engineering, Purdue University, West Lafayette
Irastorza JC, Deane RH (1974) A loading and balancing methodology for job shop control. AIIE Trans 6(4):302–307
Irdem DF, Kacar NB et al. (2008) An experimental study of an iterative simulation-optimization algorithm for production planning. In: Mason SJ, Hill R, Moench L, Rose O (eds) 2008 Winter simulation conference, Miami FL
Jackson JR (1955) Scheduling a production line to minimize maximum tardiness. University of California, Los Angeles
Jackson JR (1957) Networks of waiting lines. Opeartions Research 10(4):518–521
Johnson LA, Montgomery DC (1974) Operations research in production planning, scheduling and inventory control. Wiley, New York
Kanet JJ (1988) Load-limited order release in job shop scheduling systems. J Oper Manage 7:413–422
Karmarkar US (1987) Lot sizes, lead times and in-process inventories. Manage Sci 33(3):409–418
Karmarkar US (1989) Capacity loading and release planning with work-in-progress (WIP) and lead-times. J Manufac Oper Manage 2:105–123
Karmarkar US (1993) Manufacturing lead-times, order release and capacity loading. In: Graves SC, Rinnooy Kan AHG, Zipkin PH (eds) Logistics of production and inventory. Handbooks in operations research & management science, vol. 4. North-Holland, Amsterdam, pp 287–329
Karmarkar US, Kekre S et al. (1985a) Lotsizing in multimachine job shops. IIE Trans 13(3): 290–298
Karmarkar US, Kekre S et al. (1985b) Lot sizing and lead time performance in a manufacturing cell. Interfaces 15(2):1–9
Kekre S (1984) The effect of number of items processed at a facility on manufacturing lead time. Working paper series. University of Rochester, Rochester
Kekre S (1987) Performance of a manufacturing cell with increased product mix. IIE Trans 19(3):329–339
Kim B, Kim S (2001) Extended model for a hybrid production planning approach. International J Prod Econ 73:165–173
Kim JS, Leachman RC (1994) Decomposition method application to a large scale linear programming WIP projection model. Eur J Oper Res 74:152–160
Kim JS, Leachman RC et al. (1996) Dynamic release control policy for the semiconductor wafer fabrication lines. J Oper Res Soc 47(12):1516–1525

Kistner KP (1999) Lot sizing and queueing models: some remarks on Karmarkar's model. In: Leopold-Wildburger U, Feichtinger G, Kistner HP (eds) Modelling and Decisions in Economics: Essays in Honor of Franz Ferschl. Physica, Heidelberg, pp 173–188

Kleinrock L (1976) Queueing systems volume II: computer system applications. Wiley, New York

Koopmans T (ed) (1951) Activity analysis of production and allocation. Wiley, New York

Krämer W, Langenbach-Belz M (1976) Approximate formulae for the delay in queueing system GI/G/1. 8th International telegraphic congress. Melbourne, pp 235/1–235/8

Lambrecht MR, Chen S et al. (1996) A Lot sizing model with queueing delays: the issue of safety time. Eur J Oper Res 89:269–276

Lambrecht MR, Luyten R et al. (1984a) Protective inventories and bottlenecks in production systems. Eur J Oper Res 22:319–328

Lambrecht MR, Muckstadt JA et al. (1984b) Protective stocks in multi-stage production systems. Int J Prod Res 22:1001–1025

Land M (2004) Workload control in job shops, grasping the tap. Labyrinth, Ridderkerk

Lasserre JB, Mercé C (1990) Robust hierarchical production planning under uncertainty. Ann Oper Res 26(4):73–87

Lautenschläger M (1999) Mittelfristige Produktionsprogrammplanung mit auslastungsabhängigen Vorlaufzeiten. Peter Lang, Frankfurt am Main

Lautenschläger M, Stadtler H (1998) Modelling lead times depending on capacity utilization. Research report, Technische Universitat Darmstadt

Leachman RC (1993) Modeling techniques for automated production planning in the semiconductor industry. In: Ciriani TA, Leachman RC (eds) Optimization in industry: mathematical programming and modelling techniques in practice. Wiley, New York, pp 1–30

Leachman RC, Benson RF et al. (1996) IMPReSS: an automated production planning and delivery quotation system at Harris corporation – semiconductor sector. Interfaces 26:6–37

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

Lejeune MA, Prekopa A (2005) Approximations for and convexity of probabilistically constrained problems with random right hand sides. RUTCOR research report. Rutgers University, New Jersey

Liu L, Liu X et al. (2004) Analysis and optimization of multi-stage inventory queues. Manage Sci 50:365–380

Lu S, Ramaswamy D et al. (1994) Efficient scheduling policies to reduce mean and variance of cycle time in semiconductor plants. IEEE Trans Semicond Manufac 7:374–388

Luss H (1982) Operations research and capacity expansion problems: a survey. Oper Res 30(5):907–947

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

Manne AS (1960) On the job-shop scheduling problem. Oper Res 8(2):219–223

Medhi J (1991) Stochastic models in queuing theory. Academic, Amsterdam

Merchant DK, Nemhauser GL (1978a) A model and an algorithm for the dynamic traffic assignment problems. Transp Sci 12(3):183–199

Merchant DK, Nemhauser GL (1978b) Optimality conditions for a dynamic traffic assignment model. Transp Sci 12(3):200–207

Missbauer H (1997) Order release and sequence-dependent setup times. Int J Prod Econ 49:131–143

Missbauer H (1998) Bestandsregelung als Basis für eine Neugestaltung von PPS-Systemen. Physica, Heidelberg

Missbauer H (1999) Die Implikationen durchlauforientierter Losgrößenbildung für die Komplexität der Produktionsplanung und –steuerung. Zeitschrift für Betriebswirtschaft 69(2): 245–265

Missbauer H (2002a) Aggregate order release planning for time-varying demand. Int J Prod Res 40:688–718

Missbauer H (2002b) Lot sizing in workload control systems. Prod Plan Control 13:649–664

Missbauer H (2009) Models of the transient behaviour of production units to optimize the aggregate material flow. Int J Prod Econ 118(2):387–397
Missbauer H (forthcoming) Order release planning with clearing functions: a queueing-theoretical analysis of the clearing function concept. Int J Prod Econ
Missbauer H, Hauber W et al. (forthcoming). Developing a computerized scheduling system for the steelmaking - continuous casting process. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning in the extended enterprise: a state of the art handbook. Springer, New York
Modigliani F, Hohn FE (1955) Production planning over time and the nature of the expectation and planning horizon. Econometrica 23(1):46–66
Neuts MF (1981) Matrix-geometric solutions in stochastic models. Johns Hopkins University Press, Baltimore
Nyhuis P, Wiendahl HP (2003) Logistische Kennlinien. Springer, Berlin
Orcun S, Uzsoy R et al. (2006) Using system dynamics simulations to compare capacity models for production planning. Winter simulation conference. Monterey, CA
Orlicky J (1975) Material requirements planning: the new way of life in production and inventory management. McGraw-Hill, New York
Pahl J, Voss S et al. (2005) Production planning with load dependent lead times. 4OR 3:257–302
Parker RG (1995) Deterministic scheduling theory. Chapman and Hall, London
Parrish SH (1987) Extensions to a model for tactical planning in a job shop environment. Operations Research Center. Massachusetts Institute of Technology, Cambridge, MA
Peeta S, Ziliaskopoulos AK (2001) Foundations of dynamic traffic assignment: the past, the present and the future. Network Spatial Econ 1(3–4):233–265
Perona M, Portioli A (1998) The impact of parameters setting in load oriented manufacturing control. Int J Prod Econ 55(133–142)
Peters RJ, Boskma K et al. (1977) Stochastic programming in production planning: a case with non-simple recourse. Statistica Neerlandica 31:113–126
Philipoom RR, Fry TD (1992) Capacity based order review/release strategies to improve manufacturing performance. Int J Prod Res 30:2559–2572
Pinedo M (1995) Scheduling theory, algorithms, and systems. Prentice-Hall, New Jersey
Pinedo M, Chao X (2005) Planning and scheduling in manufacturing and services. Springer, New York
Powell SG, Schultz KL (2004) Throughput in serial lines with state-dependent behaviour. Manage Sci 50(8):1095–1105
Prekopa A (1993) Programming under probabilistic constraint and maximizing a probability under constraints. Center for operations Research, Rutgers University, New Brunswick
Rakes TR, Franz LS et al. (1984) Aggregate production planning using chance-constrained goal programming. Int J Prod Res 22(4):673–684
Riaño G (2003) Transient behavior of stochastic networks: application to production planning with load-dependent lead times. School of Industrial and Systems Engineering. Georgia Institute of Technology, Atlanta
Riaño G, Hackman S et al. (2006) Transient behavior of queueing networks. School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta
Riaño G, Serfozo R et al. (2003) Benchmarking of a stochastic production planning model in a simulation testbed. Winter simulation conference
Schneeweiß C (2003) Distributed decision making. Springer, Berlin
Selçuk B (2007) Dynamic performance of hierarchical planning systems: modeling and evaluation with dynamic planned lead times. Technische Universiteit Eindhoven, Eindhoven
Selçuk B, Fransoo JC et al. (2007) Work in process clearing in supply chain operations planning. IIE Trans 40:206–220
Sengupta JK (1972) Decision rules in stochastic programming under dynamic economic models. Swed J Econ 74:370–389
Sengupta JK, Portillo-Campbell JH (1973) A reliability programming approach to production planning. Int Stat Rev 41:115–127

Singhal J, Singhal K (2007) Holt, Modigliani, Muth and Simon's work and its role in the renaissance and and evolution of operations management. J Oper Manage 25:300–309
Smith SF (1993) Knowledge-based production management: approaches, results and prospects. Prod Plan Control 3(4):350–380
Spearman ML (1991) An analytic congestion model for closed production systems with IFR processing times. Manage Sci 37(8):1015–1029
Spearman ML, Woodruff DL et al. (1990) CONWIP: a pull alternative to Kanban. Int J Prod Res 28(5):879–894
Spitter JM, de Kok AG et al. (2005a) Timing production in LP models in a rolling schedule. Int J Prod Econ 93–94:319–329
Spitter JM, Hurkens CAJ et al. (2005b) Linear programming models with planned lead times for supply chain operations planning. Eur J Oper Res 163:706–720
Srinivasan A, Carey M et al. (1988) Resource pricing and aggregate scheduling in manufacturing systems. Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh
Stadtler H (1996) Hierarchische Produktionsplanung. Handwörterbuch der Produktionswirtschaft. Schäffer-Poeschel, Stuttgart, pp 631–641
Stadtler H, Kilger C (eds) (2008) Supply chain management and advanced planning: concepts, models, software and case studies. Springer-Verlag, Berlin
Stange K (1964) Die Anlauflösung für den einfachen exponentiellen Bedienungskanal (mit beliebig vielen Warteplätzen), der für t=0 leer ist. Unternehmenforschung 8:1–24
Stevenson M, Hendry LC (2006) Aggregate load-oriented workload control: a review and a re-classification of a key approach. Int J Prod Econ 104(2):676–693
Tang L, Liu J et al. (2001) A review of planning and scheduling systems and methods for integrated steel production. Eur J Oper Res 133:1–20
Tardif V, Spearman ML (1997) Diagnostic scheduling in finite-capacity production environments. Comput Ind Eng 32:867–878
Tatsiopoulos IP, Kingsman BP (1983) Lead time management. Eur J Oper Res 14:351–358
Tijms HC (1994) Stochastic models: an algorithmic approach. Wiley, New York
Uzsoy R, Lee CY et al. (1994) A review of production planning and scheduling models in the semiconductor industry part II: shop-floor control. IIE Trans Scheduling Logistics 26:44–55
Van Ooijen HPG (1996) Load-based work-order release and its effectiveness on delivery performance improvement. Eindhoven University of Technology, Eindhoven
Van Ooijen HPG, Bertrand JWM (2003) The effects of a simple arrival rate control policy on throughput and work-in-process in production systems with workload dependent processing rates. Int J Prod Econ 85(1):61–68
Vaughan TS (2006) Lot size effects on process lead time, lead time demand, and safety stock. Int J Prod Econ 100:1–9
Vepsalainen AP, Morton TE (1987) Priority rules for job shops with weighted tardiness costs. Manage Sci 33(8):1035–1047
Vepsalainen AP, Morton TE (1988) Improving local priority rules with global lead-time estimates: a simulation study. J Manufac Oper Manage 1:102–118
Vollmann TE, Berry WL et al. (1988) Manufacturing planning and control systems. Richard D. Irwin, Boston
Vollmann TE, Berry WL et al. (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
Wagner HM, Whitin TM (1958) Dynamic version of the economic lot size model dynamic version of the economic lot size model. Manage Sci 5:89–96
Wiendahl HP (1995) Load oriented manufacturing control. Springer, Heidelberg
Wight O (1983) MRPII: unlocking America's productivity potential. Oliver Wight, Williston
Wijngaard J, Wortmann JC (1985) MRP and inventories. Eur J Oper Res 20:281–293
Yano CA, Carlson RC (1988) Safety stocks for assembly systems with fixed production intervals. J Manufac Oper Manage 1:182–201

Zäpfel G, Missbauer H (1993a) Production planning and control (PPC) systems including load-oriented order release – problems and research perspectives. Int J Prod Econ 30:107–122
Zäpfel G, Missbauer H (1993b) New concepts for production planning and control. Eur J Oper Res 67:297–320
Zäpfel G, Missbauer H et al. (1992) PPS-Systeme mit belastungs orientierter Auftragsfreigabe – Operationscharakteristika und Möglichkeiten zur Weiterentwicklung. Zeitschrift für Betriebswirtschaft 62:897–919
Zipkin PH (1986) Models for design and control of stochastic, multi-item batch production systems. Oper Res 34(1):91–104
Zipkin PH (1997) Foundations of inventory management. Irwin, Burr Ridge
Zweben M, Fox M (eds) (1994) Intelligent scheduling systems. Morgan Kaufman, San Francisco