Chapter 7
Univariate Clearing Functions

In this chapter, we introduce the concept of the clearing function (CF), a metamodel of a production resource that relates the expected output of a resource to some measure of the work available to it in the planning period. We focus on clearing functions with a single state variable and examine a variety of functional forms that have been proposed in the production and traffic literature. We then formulate release planning models using these functions and show that while single-product models yield tractable convex optimization problems, the presence of multiple products competing for capacity at a shared resource creates significant difficulties. The allocated clearing function formulation is presented to address these issues and shown to yield more informative dual prices for resource capacity than conventional LP models.

7.1 Preliminaries

The models in the previous two chapters anticipate the performance of the production units using exogenous, workload-independent lead times that are assumed to remain valid as long as a maximum capacity loading is not exceeded. These lead time estimates may take different forms based on how capacity is consumed during the lead time, as discussed in Chap. 5, and can be specific to individual planning periods as discussed in Chap. 6.

The combination of fixed, exogenous planned lead times with a maximum capacity limit as an anticipation function yields computationally tractable linear programming (LP) models as long as lot sizing is not a consideration; the presence of lot sizing requires the introduction of integer variables and yields considerably more challenging models (Pochet and Wolsey 2006). As long as the production unit operates at approximately constant utilization over time, historical data can be used to estimate planned lead times that are consistent with observed cycle times, for example by setting the planned lead times to a specified fractile of the observed cycle

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_7

time distribution. However, if the resource utilization level, the product mix, or both vary over time, the distribution of the cycle time will also change over time. This, in turn, may cause the cycle times observed on the shop floor to deviate significantly from the lead times used in the planning models, adversely affecting the performance of the production units trying to execute these plans.

In contrast to these LP models where the output of the system is determined by the combination of planned lead times and a maximum capacity loading, the models in this chapter express the expected output of the production unit in a planning period as a function of the workload available to the resource for processing in that period. Models of this type have arisen in the context of queueing systems, in the management of traffic networks and as representations of particular production control policies. We shall refer to models of this type as *clearing functions*, following the terminology of Karmarkar (1989).

We define a clearing function as a functional relationship that specifies the expected output X_t of a production resource in a planning period t of duration Δ as

$$X_t = f\left(\Delta, \Omega_t\right) \tag{7.1}$$

where Ω_t denotes a set of state variables that collectively describe the amount of work available to the resource in period t. The specific set of state variables to include in the set Ω_t is not immediately obvious. From a queueing perspective, the state of the resource at time t potentially depends on the entire past history of the relevant stochastic processes (interarrival times, service times, machine failures, setups, number of available machines, etc.) up to that instant in time. It is also apparent that the clearing function must depend on the length Δ of the planning period for which it is being constructed. Finally, the amount of work available to the resource and the distribution of its arrival over time depend on the model used by the planning level to determine releases over time. In queueing terms, the release decisions made by the planning level affect both the mean interarrival time of orders to the resource and its variance.

The purpose of the clearing function is to represent the behavior of the resource to an acceptable degree of accuracy while still yielding tractable optimization models for the planning problem. The extremely high dimensionality and complex functional forms required by general methods, such as queueing approaches considering the entire history of the process or a large portion of it, make it very difficult to obtain clearing functions leading to tractable optimization models. Even simple functional forms for clearing functions can yield non-convex optimization models. Hence most clearing functions proposed to date have used a single state variable; we shall see that even in this case formulations involving multiple products can become challenging. In this chapter, we discuss various single-variable clearing functions, the difficulties that arise when multiple products compete for capacity at a resource, and solutions to these difficulties. We also show that planning models using clearing functions can produce meaningful dual prices for resources at any level of utilization, which is not the case for the models discussed in Chaps. 5 and 6.

7.2 Single-Variable Clearing Functions

7.2.1 Average WIP-Based Clearing Functions

This family of clearing functions, the motivation for which was sketched in Sect. 2.2, uses the set of state variables $\Omega_t = \{\bar{W}_t\}$, where $\bar{W}_t$ denotes the time-average WIP level, measured in number of units or lots, at the production resource over the planning period t. Specifically, if planning period t spans the time interval $(t-\Delta, t]$ and $W(t)$ denotes the amount of WIP at the resource at time t, we have

$$\bar{W}_t = \frac{1}{\Delta}\int_{t-\Delta}^{t} W(\tau)\,d\tau \tag{7.2}$$

The advantage of $\bar{W}_t$ as a workload metric is its straightforward relation to the well-known steady-state analyses of queues such as the *M/G/1* and *G/G/1* (Buzacott and Shanthikumar 1993; Curry and Feldman 2000), from which exact or approximate expressions relating the expected WIP, expected cycle time and utilization can be derived. As discussed in Sect. 2.2, the expected WIP level of the *G/G/1* queue in steady state is given by

$$\bar{W} = \frac{T}{t_a} = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u}{1-u}\right)\frac{t_e}{t_a} + \frac{t_e}{t_a} = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u^2}{1-u}\right) + u \tag{7.3}$$

dropping the time subscript since this is a steady-state relation. Solving for u in terms of $\bar{W}$ yields a quadratic equation in $\bar{W}$ whose nonnegative solution is

$$u = \frac{-(\bar{W}+1) + \sqrt{(\bar{W}+1)^2 + 4(\psi-1)\bar{W}}}{2(\psi-1)} \tag{7.4}$$

where $\psi = (c_a^2 + c_e^2)/2$; recall from Sect. 2.2 that $\psi = 1$ represents the special case of the *M/M/1* queue. Intuitively, the higher the average WIP level $\bar{W}$ at the resource, the lower the probability $(1 - u)$ that the resource will be idle due to lack of work; hence maintaining a planned average throughput rate of X in a planning period requires maintaining a certain average WIP level at the resource. The average utilization can be interpreted as the fraction of the planning period during which the resource will be producing usable output. Thus the expected number of units produced over the planning period is given by:

$$X_t = \frac{u\Delta}{t_e} = \frac{\Delta}{t_e}\left[\frac{-(\bar{W}+1) + \sqrt{(\bar{W}+1)^2 + 4(\psi-1)\bar{W}}}{2(\psi-1)}\right] \tag{7.5}$$

Incorporating this state variable into optimization models requires some additional considerations. At the point in time the optimization model is solved to determine releases for the next T periods, the average WIP values $\bar{W}_t$ are not known with certainty; they are in fact random variables whose distribution is determined by the release decisions made by the planning model. Hence the state variables W_t representing WIP in the optimization models actually represent the planned state of the resource at the end of period t and do not capture the evolution of the WIP level throughout the planning period. However, many different WIP trajectories $W(t)$ may give the same beginning and ending WIP levels W_{t-1} and W_t. Optimization models using this type of clearing function must estimate the planned value of $\bar{W}_t$ for a given planning period t using the planned values of W_{t-1} and W_t. The most obvious approach is to use the arithmetic average to obtain

$$\bar{W}_t = \frac{\left(W_{t-1} + W_t\right)}{2} \tag{7.6}$$

However, this has implications for the behavior of the resulting optimization models. Note that if W_{t-1} is increased by a certain amount in (7.6) and W_t reduced by the same amount, $\bar{W}_t$ remains unchanged. Depending on the structure of the optimization model, this can lead to oscillating WIP levels at the period boundaries due to the presence of alternative optimal solutions, which is undesirable (Missbauer (1998): 413 ff.).

Given their origin in steady-state queueing analysis, clearing functions of this type are more appropriate for longer planning periods, where the transient behavior of the resource at the start of the period due to changes in releases can safely be neglected. Note that it is possible to have $X_t \geq \bar{W}_t$ using a clearing function of this type; at low utilization levels, the average queue length will be very small, while the total output will be approximately equal to the number of arrivals during the period.

Although it is not explicitly stated as such, the practical worst case model of production lines given in Chap. 7 of Hopp and Spearman (2008) also represents an average WIP-based clearing function. This model considers a balanced serial production line operating under the CONWIP policy discussed in Chap. 4. They define the system state as a vector whose components represent the number of jobs in front of each machine in the line. Assuming all such states to be equally likely, they note that for a total WIP level of w jobs in the line, a new job entering the system will see on average

$$W_i = \frac{\left(w-1\right)}{N} \tag{7.7}$$

jobs ahead of it at each of the N machines in the system, implying an average cycle time of

$$T = T_0 + \frac{w-1}{r_b} \tag{7.8}$$

where T_0 denotes the raw processing time of the line, the average time in system a job will encounter if it enters an empty line, and r_b the processing rate of the bottleneck machine. Substituting (7.8) into Little's Law yields an average throughput rate of

$$X = \frac{wr_b}{r_b T_0 + w - 1} \tag{7.9}$$

Since in a CONWIP system the average WIP level will be equal to the total WIP level w permitted in the system, this represents an average WIP-based clearing function that can be shown to be concave and monotonically non-decreasing in the average WIP level w. The assumption of equally likely system states is exact only for a balanced line with a single exponential server at each stage, but provides a WIP level that is unlikely to be exceeded in systems with more general structures.

7.2.2 Initial WIP-Based Clearing Functions

This family of clearing functions assumes that the expected output of the resource in a planning period is determined solely by the amount of work available to it at the start of the planning period; work arriving during the period will have no effect on expected output. Hence the set of state variables considered in each period t is $\Omega_t = \{W_{t-1}\}$. Under this model either the probability of new work arriving during the planning period is negligible, the scheduling policy only allows work to be released at the start of a period (which coincides with the end of the previous one), or the planning interval is sufficiently short that work available at the start of the period will fully occupy the resource until the next period.

Clearing functions of this type have been discussed extensively in the context of traffic assignment problems (Dafermos and Sparrow 1969; Carey 1987; Peeta and Ziliaskopoulos 2001) where they are used to model the behavior of a section of highway in a given time period. In these networks, which bear considerable similarity to those studied in this volume, a traffic system is modeled as a network with node set N and directed arc set A. The arcs $(i,j) \in A$ correspond to specific segments of roadway whose starting and ending points are represented by nodes $i, j \in N$, respectively. The amount of traffic $X_{ij}(t)$ that can exit the arc (i,j) over a planning period t is expressed as a concave, non-decreasing function $g_{ij}(W_{ij}(t-1))$ of the amount of traffic $W_{ij}(t-1)$ present on the arc at the start of the period. These exit functions are used in discrete-time optimization models very similar to those developed later in this chapter.

The exit functions used in the dynamic traffic assignment work are derived from flow-density functions, which are discussed in detail in Carey and Bowers (2012). The basic resource considered in these models, analogous to the machine or workcenter in production units, is a segment of road whose characteristics such as width, surface quality, visibility, and signage are assumed to be known. For ease of exposi-

tion we shall assume the road segment to be of unit length, and will drop the time subscript to discuss a generic time period, as in the discussion of steady-state clearing functions in Sect. 7.2.1. The progress of individual vehicles along the road segment is represented as a continuous flow, in much the same manner as the LP models of Chap. 5 treat the processing of discrete orders at the production resources. The traffic density k represents the number of vehicles occupying the road segment of unit length being considered. This quantity is analogous to the average WIP $\overline{W}_t$ or workload in production contexts. The flow rate q, the number of vehicles passing a particular point on the road per unit time, is analogous to the throughput rate X of a workcenter or production resource. Hence the exit function captures the rate at which vehicles pass the end point of the road segment, either entering another segment or exiting the system. The space mean speed v of the traffic along the unit road segment is given by the length of the road segment divided by the average time to traverse it. The relation between flow rate q, speed v, and traffic density k is thus

$$q = kv \tag{7.10}$$

Noting that $v = 1/T$, where T denotes the average time to traverse the road segment, we obtain

$$q = \frac{k}{T} \tag{7.11}$$

which can be rewritten as

$$k = qT \tag{7.12}$$

Replacing each term with its counterpart in the production context (k with $\overline{W}$ and q with X) and noting that the interpretation of T as the average time to traverse the system under consideration is the same in both traffic and production contexts, we recover Little's Law (Hopp and Spearman 2008):

$$W = XT \tag{7.13}$$

Flow-density functions $q = f(k)$ are intended to be empirical relations whose parameters are estimated from appropriately collected data. However, most flow-density functions $f(.)$ used in traffic research have been derived using a limited set of parameters:

– The free-flow velocity V_0 of the road segment, representing the flow of traffic at very low density, analogous to the raw process time T_0 discussed in the previous section. Since by (7.10) the average velocity $v = q/k = f(k)/k$, we have

$$V_0 = \lim_{k \to 0^+} \frac{f(k)}{k} = \left. \frac{\partial f}{\partial k} \right|_{k=0} \tag{7.14}$$

– The jam density k_j, the density at which $v = q = 0$, i.e., traffic comes to a stop.

– The wave speed at jam density c_j, the rate at which flow decreases as density increases to the jam density k_j, given by

$$c_j = \lim_{k \to k_j} \frac{df(k)}{dk} \tag{7.15}$$

– The maximum flow rate q_c. The density at which the maximum flow rate occurs is referred to as the critical density k_c, analogous to the critical WIP concept of Hopp and Spearman (2008).

Carey and Bowers (2012) propose several desirable properties for a flow-density function. These include unimodality, appropriate finite values of the free-flow speed V_0, jam density k_j, and the ratio k_j/k_c, as well as an appropriate negative value of c_j and the possibility of convexity as $k \to k_j$. A generic flow-density function $f(k)$ satisfying these conditions would appear as shown in Fig. 7.1.

Production systems research has generally assumed an infinite jam density $k_j = \infty$, under the assumption that as the work available to a queueing system in a planning period increases its output rate X will eventually level off at $1/t_e$, but will never decrease. In environments where jobs do not interfere with each other through sequence-dependent setup times or scheduling policies, this assumption appears reasonable. Hence most clearing functions proposed by production system researchers have taken the form of monotonically non-decreasing concave functions that asymptotically approach the maximum production rate as workload or WIP approach infinity. Clearing functions for environments where this assumption is not valid, such as those with significant sequence-dependent setup times, are discussed in the next two chapters. Clearing functions that decrease beyond a certain WIP level like the flow-density function in Fig. 7.1, due to e.g., reduced worker efficiency when workload is too high or by excessive material shuffling which reduces capacity, are rare in the literature (Van Ooijen and Bertrand 2003).

While a wide range of flow-density functions have been discussed in the traffic research community, we will use two examples to illustrate the types of models

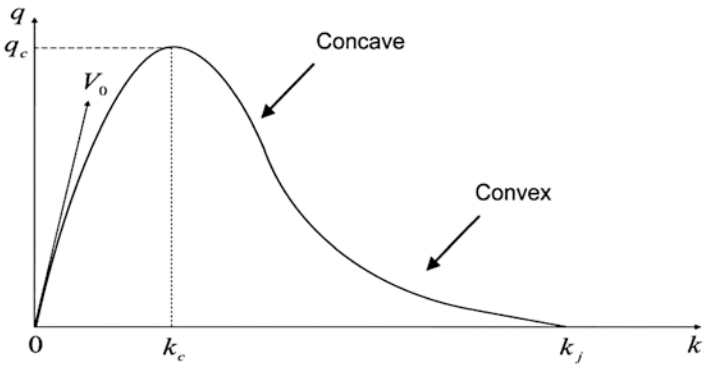

Fig. 7.1 A generic flow-density function (Carey and Bowers 2012)

considered. The output function proposed by Newell (1961) and Franklin (1961) takes the form

$$f(k) = kV_0\left(1 - \exp\left[\frac{|c_j|}{V_0}\left(1 - \frac{k_j}{k}\right)\right]\right) \tag{7.16}$$

Carey and Bowers (2012) note that this flow-density function satisfies more of the desirable properties they propose than any other function; however, it is concave everywhere, not admitting convexity as the jam density is approached. They also point out that the function is defined by three parameters (k_j, V_0, and c_j) that give the behavior of the function at the origin and at jam density unduly high influence on its overall shape. This function appears to have motivated the clearing function of Srinivasan et al. (1988) discussed below. Another class of flow-density functions proposed by Van Aerde and Rakha (1995) takes the form

$$f(k) = \frac{1 - (c_1 - c_3V_0)k - \sqrt{\left[(c_1 + c_3V_0) - 1\right]^2 + 4c_3c_2k^2}}{2c_3k} \tag{7.17}$$

where c_1, c_2, and c_3 are constants computed from V_0, k_j, q_c, and v_c, where the latter denotes the average speed at critical density q_c. The resemblance to (7.4) is striking.

7.2.3 Workload-Based Clearing Functions

The discrete-time nature of production planning models creates difficulties for average WIP-based clearing functions due to the fact that multiple combinations of values for W_t and W_{t-1} can yield the same $\bar{W}_t$ value for any period t. Initial WIP-based clearing functions assume that the expected output X_t of the resource in period t cannot exceed the initial WIP W_{t-1} available at the start of the period, ignoring the possibility that work released during the period might be completed during the period. Workload-based clearing functions address this issue by using a state variable Λ_t that represents the total amount of work made available to the resource during period t, given by

$$\Lambda_t = W_{t-1} + R_t \tag{7.18}$$

where W_{t-1} denotes the amount of WIP carried over from the previous period $t - 1$ and R_t the amount of work released to the resource during period t. Clearing functions of this form must have

$$\frac{\partial f}{\partial \Lambda_t} \leq 1 \tag{7.19}$$

for all $\Lambda_t \geq 0$, implying that the resource can never convert more material into output in a period than becomes available to it over the period.

Missbauer (2002) proposes a clearing function of this form for a resource that can be represented as an *M/G/1* queue in steady state. We present here the same development for a *G/G/1* queue. Recall from (7.5) that the expected throughput of a *G/G/1* queue in steady state can be approximated as

$$X_t = \frac{u\Delta}{t_e} = \frac{\Delta}{t_e}\left[\frac{-\left(\bar{W}+1\right)+\sqrt{\left(\bar{W}+1\right)^2+4\left(\psi-1\right)\bar{W}}}{2\left(\psi-1\right)}\right] \tag{7.20}$$

Now consider a *G/G/1* queue in steady state where at the start of some planning period t there are W_{t-1} units of work remaining on hand from the previous period and R_t units are released into the production unit. Recalling that the workload $\Lambda_t = W_{t-1} + R_t$, we have $W_{t-1} = \Lambda_t - R_t$. Since the queue is assumed to be in steady state, we must have $X_t = R_t$ and $W_{t-1} = \bar{W}_t$. Substituting $\bar{W}_t = \Lambda_t - X_t$ into (7.20) and solving for X_t, we obtain

$$X_t = \frac{\Delta}{t_e}\left[\frac{\left(\Delta+t_e\left(1+\Lambda_t\right)\right)-\sqrt{\left[\Delta+\left(1+\Lambda_t\right)\right]^2-4\Lambda_t t_e\left[\Delta+t_e\left(\psi-1\right)\right]}}{2\left[\Delta-t_e\left(\psi-1\right)\right]}\right] \tag{7.21}$$

The basic form of this expression is quite similar to that derived for the average WIP case in (7.20); most notably, it retains the concave saturating form and guarantees that $X_t \leq \Lambda_t$. Its drawback is the assumption of steady state, which is not generally valid under the conditions of time-varying demand and finite period length under which we wish to use the release planning models we study. Again, we note in passing the similarity to (7.17).

7.2.4 The Constant Cycle Time Clearing Function

Graves (1986) proposes a discrete-time model of a production resource whose expected output X_t in period t is given by the clearing function

$$X_t = \alpha W_{t-1} \tag{7.22}$$

where W_{t-1} denotes the amount of WIP available to the resource at the start of period t, i.e., the end of period $t-1$. Since Graves assumes that work can only arrive at or depart from the resource at the start of a planning period, this can also be viewed as a workload-based clearing function in our terminology. The resource will always process a fraction α of the WIP W_{t-1} available to it at the start of the period, no matter how large W_{t-1} may be. Equivalently, the model assumes that the resource is managed to maintain an average cycle time of $1/\alpha$ periods; as the amount of avail-

able work W_{t-1} increases, the resource can work faster. Hence this linear clearing function is best viewed as describing the behavior of the production resource under a specified production control policy, where the processing rate can be varied to maintain the planned lead time of $1/\alpha$ periods. The clearing function will, naturally, only be valid over the range of operating conditions that satisfy this condition.

The author uses clearing functions of this type to analyze the performance of a job shop by computing the mean and variance of performance measures such as throughput, queue length, and backlog. In particular, he examines the tradeoff between production smoothing (which requires long planned lead times and hence low values of α) and reducing cycle times and WIP levels (which requires high values of α) by simulating a job shop environment. The author uses this model in several subsequent papers to examine the issue of setting safety stocks in such systems (Graves 1988), planning in multistage production-inventory systems (Graves et al. 1998), and setting planned lead times in make-to-order systems (Teo et al. 2011; Teo et al. 2012). Parrish (1987) extends the model to a network of workcenters in a transient regime.

7.2.5 *Empirically Based Single-Variable Clearing Functions*

These are functional forms that have been used to fit clearing functions empirically to data obtained from either industrial data or simulation. One or another of the clearing function families discussed above is used to postulate a basic functional form whose parameters are then fitted to empirical data gathered from either direct observation of the production unit or, more frequently, a simulation model.

Karmarkar (1989) proposes a workload-based clearing function of the form

$$X_t = \min\left\{\Lambda_t, \frac{K_1\Lambda_t}{K_2 + \Lambda_t}\right\} \tag{7.23}$$

motivated by the clearing function for an *M/M/1* queue. Here K_1 represents the maximum expected output of the resource assuming unlimited workload and K_2 a user-determined parameter governing the curvature of the clearing function. In general, K_2 is increasing in the amount of variability in the system as described by the coefficients of variation of the service times and interarrival times. The clearing function is given as the minimum of two quantities to ensure that output does not exceed the total workload available to the resource; this can also be achieved by selecting the value of K_2 such that $\left.\frac{\partial X_t}{\partial \Lambda_t}\right|_{\Lambda_t=0} = 1$. This function is concave and monotonically non-decreasing, with $\lim_{\Lambda_t \to \infty} X_t = K_1$.

The functional form $X_t = K_1\Lambda_t/(K_2 + \Lambda_t)$ in (7.23) originates from the functional relationship between average WIP (in contrast to the workload Λ_t) and output; therefore, it can exceed the available workload in period t. Missbauer (2002) shows that for the *M/G/1* model in equilibrium, the expected output X_t and expected load Λ_t of a workcenter are related as follows:

$$X_t = \frac{1}{2}\left(K_1 + K_2 + \Lambda_t - \sqrt{K_1^{\,2} + 2K_1K_2 + K_2^{\,2} - 2K_1\Lambda_t + 2K_2\Lambda_t + \Lambda_t^{\,2}}\right) \quad (7.24)$$

with K_1 the maximum expected output of the resource (capacity) as above, $K_2 = \frac{\sigma^2}{2t_e} + \frac{t_e}{2}$ and σ^2 the variance of the service times. This function is analogous to (7.21) and can be parameterized using empirical or simulated data.

Srinivasan et al. (1988) suggest an initial WIP-based clearing function similar to the flow-density function (7.16), given by

$$X_t = K_1\left[1 - \exp\left(-K_2 W_{t-1}\right)\right] \quad (7.25)$$

Here K_1 again represents the maximum expected output of the resource with unlimited WIP, and K_2 a user-defined parameter governing the curvature of the clearing function. Once again we have $\lim_{\Lambda_t \to \infty} X_t = K_1$.

Concave, saturating functional forms of clearing functions derived from queueing models usually approach their limit (the maximum possible expected output) asymptotically because the underlying assumptions of renewal processes usually allow arbitrarily long interarrival and service times. In reality, this is often not the case since the order release system will try to prevent very long interarrival times and service times can be controlled by lot sizing. Nyhuis and Wiendahl (2009) suggest defining threshold values $\bar{W}^{\mathrm{u}}$ and $\bar{W}^{\mathrm{o}}$ with $\bar{W}^{\mathrm{u}} < \bar{W}^{\mathrm{o}}$ for the average WIP $\bar{W}$ where for $\bar{W} < \bar{W}^{\mathrm{u}}$ output is proportional to $\bar{W}$, as in the "Best Case" clearing function of Hopp and Spearman (2008), and for $\bar{W} > \bar{W}^{\mathrm{o}}$ the workcenter is fully utilized. Appropriate functional forms are derived. In order to apply this logic to a period-based clearing function with the workload Λ_t as state variable, threshold values Λ_t^{u} and Λ_t^{o} with $\Lambda_t^{\mathrm{u}} \le \Lambda_t^{\mathrm{o}}$ are defined for the workload, leading to different clearing functions for different regimes of operation such that

$$X_t = \begin{cases} \Lambda_t, & 0 < \Lambda_t \le \Lambda_t^{\mathrm{u}} \\ f\left(\Lambda_t\right), & \Lambda_t^{\mathrm{u}} \le \Lambda_t \le \Lambda_t^{\mathrm{o}}. \\ C_t, & \Lambda_t \ge \Lambda_t^{\mathrm{o}} \end{cases} \quad (7.26)$$

In this case, the problem of estimating the clearing function is essentially that of estimating its deviation from the ideal shape $X_t = \mathrm{Min}\,(\Lambda_t; C_t)$.

7.3 Piecewise Linear Single-Variable Clearing Functions

Many authors using single-variable clearing functions in optimization models have chosen to approximate the concave clearing function by outer linearization. This approach has several benefits: it allows the overall production planning model to take the form of a linear program, which is computationally tractable and scalable. In addition, piecewise linearization of a univariate clearing function proves extremely useful in the development of clearing function models for multiple-item systems. We shall present the ideas in this section using the workload-based clearing function as our vehicle, but the basic issues are relevant to all concave single-variable clearing functions.

It is well known in convex analysis, as a consequence of the Fenchel-Young Theorem (that any convex region can be represented as the supremum of its affine minorants) (Boyd and Vandenberghe 2009), that any convex function can be approximated to any desired degree of accuracy by the convex hull of a set of affine functions of the form

$$f^q\left(\Lambda_t\right) = \alpha^q \Lambda_t + \beta^q, \quad q = 1,\ldots,Q \tag{7.27}$$

In order to reflect the concavity of the original clearing function $f(\Lambda_t)$, we assume that the segments have slopes such that $\alpha^1 \geq 1 > \alpha^2 > \ldots > \alpha^Q = 0$, and intercepts $0 = \beta^1 < \beta^2 < \ldots < \beta^Q$. The intercept β^Q of the final segment represents the maximum possible expected output from the production unit in a time period, while the slope α^1 of the first segment is bounded above by 1, since even at very low workloads there may be a nonzero probability of some work remaining incomplete at the end of the period if, for example, a large fraction of the workload arrives very late in the period.

Given a concave clearing function of whatever specific functional form, the problem of determining the best piecewise linear approximation can be formulated as an optimization problem in several different ways. We shall describe one such formulation described by Turkseven (2005), which we shall refer to as the trapezoidal formulation, to illustrate the basic approach. Imamoto and Tang (2008) present an alternative formulation that minimizes the maximum error of the piecewise linear approximation for a given number of segments.

For illustrative purposes, we shall consider the problem of obtaining the best piecewise linear approximation to a concave non-decreasing clearing function $f(\Lambda_t)$ using three linear segments of the form (7.27) as seen in Fig. 7.2. Let t_q, $q = 1, \ldots, Q$ denote the value of Λ_t at which segments q and q+1 intersect, and a_q, $q = 1, \ldots Q$ the value of Λ_t at which the qth linear segment is tangent to the concave clearing function. Additionally we define $t_0 = 0$ and $t_{Q+1} = \Lambda_{\max}$, an upper limit on the workload considered. For given values of α^q and β^q straightforward geometry gives

$$a_q = \beta_q \frac{\left(\alpha_{q+1} - \alpha_q\right)}{\left(\beta_{q+1} - \beta_q\right)} + \alpha_q \tag{7.28}$$

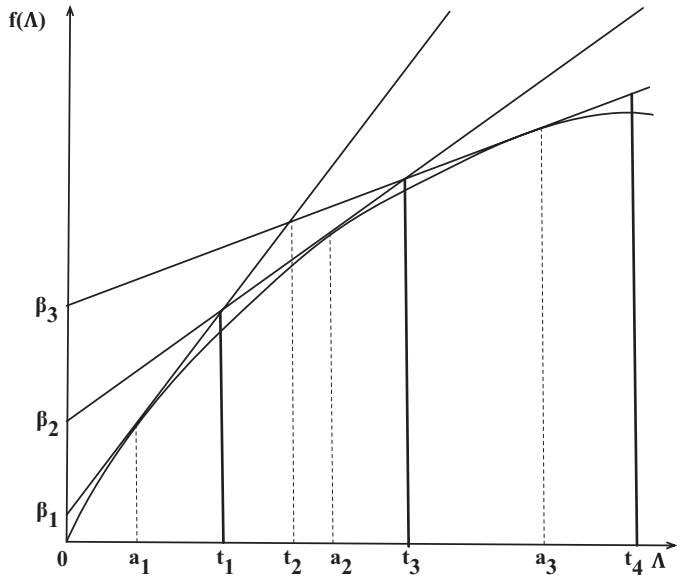

Fig. 7.2 Illustration of trapezoid formulation for piecewise linearization of a concave clearing function

$$t_q = \frac{\alpha_{q+1} - \alpha_q}{\beta_{q+1} - \beta_q} \tag{7.29}$$

The decision variables in the optimization formulation are the slopes α^q and intercepts β^q of the linear segments $q = 1, ..., Q$. The objective function to be minimized is given by the difference in the areas under the convex clearing function and its piecewise linear approximation, which is equivalent to minimizing the area under the piecewise linear approximation when the segments q are constrained to be tangent to the original clearing function. For Q linear segments, the area under the piecewise linear approximation will consist of Q trapezoids, with the area of the trapezoid formed by segments q and $q + 1$ given by

$$A_q = \beta^q + \frac{1}{2}\alpha^q \left(t_{q-1} + t_q\right) \tag{7.30}$$

The optimization model can then be written as

$$\min \sum_{q=1}^{Q} A_q \tag{7.31}$$

subject to

$$\alpha^q t_q + \beta^q = f\left(t_q\right), \quad q = 1, \ldots, Q \tag{7.32}$$

$$\beta^q = \left.\frac{df(\Lambda)}{d\Lambda}\right|_{\Lambda=a_q}, \quad q = 1,\ldots,Q \tag{7.33}$$

$$\alpha^q, \beta^q \geq 0, \quad q = 1,\ldots,Q \tag{7.34}$$

The fractional structure of (7.28) and (7.29) generally results in a non-convex nonlinear formulation for which a global optimum is hard to obtain in reasonable CPU times. Imamoto and Tang (2008) provide an exact recursive algorithm for their minimax formulation, while Turkseven (2005) proposes an alternative heuristic for the trapezoid formulation. Asmundsson et al. (2009) solve the trapezoid formulation with a standard convex nonlinear solver, obtaining a local optimum that appears satisfactory in most cases, although some instances where the solver failed to converge were also encountered.

We make no claim as to the originality of the trapezoid formulation; it is one of several fairly obvious approaches to the problem, and has almost certainly been formulated before, although we have been unable to find the original reference. We provide it here for the sake of completeness. However, recent work by Gopalswamy and Uzsoy (2019) suggests that rather than fitting a nonlinear functional form to data and then piecewise linearizing this concave function, directly fitting a piecewise linear concave function to the data using convex regression (Toriello and Vielma 2012; Hannah and Dunson 2013; Gopalswamy et al. 2019) yields considerably better results.

7.4 Optimization Models for a Single Production Resource

The clearing functions presented above all represent the system state in an aggregate manner; the workload Λ_t, the initial WIP W_{t-1}, or the time-average WIP $\bar{W}_t$ are aggregated over the different products in the system, in a manner similar to that used by queueing models of multi-item systems: the mix of different items arriving randomly at the resource over time results in the effective service times following a probability distribution whose first and second moments can be used to derive a clearing function. However, any useful production planning model must determine the mix of products to be released into the system in each planning period t, requiring disaggregation if an aggregate single-variable clearing function is used. The development of clearing function models for multiple-item systems presents a number of challenges; similar issues are encountered in traffic modeling with multiple vehicle classes or origin-destination pairs (Carey 1992). These difficulties have proven to be persistent in both research areas, and merit detailed discussion since a fully satisfactory solution remains elusive.

To illustrate the issues, we first present a model of a simple single-product problem, closely following the development of Karmarkar (1989). For ease of exposition, we assume a time-stationary workload-based clearing function $f(\Lambda_t)$ and

time-stationary cost parameters. We also assume no backlogging of unmet demand is allowed; if present, it can be incorporated easily (Johnson and Montgomery 1974). We define the following notation:

Indices:

t: planning period, $t = 1,\ldots,T$. $t = 0$ will be used to denote the initial state of the system at the start of period 1, i.e., the end of period 0.

Parameters:

c: unit production cost
h: unit finished goods inventory holding cost
w: unit WIP holding cost
r: unit cost of raw materials, incurred upon release of the material to the production unit
$f(\Lambda_t)$: clearing function representing the behavior of the production unit, which we assume to be a concave monotonically non-decreasing function of Λ_t
D_t: demand in period t
I_0: amount of product in finished goods inventory at the start of period 1
W_0: amount of product in WIP at the start of period 1

Decision Variables:

X_t: output of production unit in period t, in units of product
R_t: amount of product released into production unit in period t
I_t: amount of product in finished goods inventory at the end of period t
W_t: amount of product remaining in WIP at the end of period t

In the fixed lead time models of Chap. 5, material released at the start of period t subject to a fixed lead time L emerges as finished product at the start of period $t+L$. Thus the output of the production unit is simply the time-shifted release schedule. However, in clearing function models the output of the resource in a given period t is driven only indirectly by the releases R_t. In a given period t, the resource is assumed to have $W_{t-1} \geq 0$ units of WIP remaining from the previous period. R_t units of product are released to the resource, resulting in a workload of $\Lambda_t = W_{t-1} + R_t$ units. The output of the resource during this period t is then determined by the clearing function as $X_t = f(\Lambda_t)$. These dynamics yield the following single-product clearing function (SPCF) model:

$$\min \sum_{t=1}^{T} \left[rR_t + cX_t + hI_t + wW_t \right] \tag{7.35}$$

subject to

$$W_t = W_{t-1} + R_t - X_t, \quad t = 1,\ldots,T \tag{7.36}$$

$$I_t = I_{t-1} + X_t - D_t, \quad t = 1,\ldots,T \tag{7.37}$$

$$X_t \le f\left(\Lambda_t\right), \quad t = 1,\ldots,T \tag{7.38}$$

$$R_t, X_t, I_t, W_t \ge 0, \quad t = 1,\ldots,T \tag{7.39}$$

The objective function (7.35) minimizes the sum of raw material, production, finished goods holding and WIP holding costs over the planning horizon of T periods. Constraints (7.36) are material balance equations for the WIP, and constraints (7.37) those for finished goods inventory. Constraints (7.38) limit the output in each period by the clearing function, while (7.39) ensure nonnegativity of the decision variables. Like most of the LP models discussed in Chap. 5, the SPCF model can be represented as a network flow model on a time-replicated network as shown in Fig. 7.3.

Several differences from the models of Chap. 5 are worth highlighting. First of all, no lead times appear in the formulation; the delay between material being released and its emergence as finished product capable of meeting demand is implied by the clearing function constraints (7.38). Since the argument of the clearing function depends on the WIP variables W_t, material balance constraints (7.36) are required to keep track of these variables. This distinction between WIP and finished goods inventory is intuitive, since in practice these inventories serve different purposes. Production is made possible by having sufficient WIP in the system, while finished goods inventory, represented by the I_t variables, allows inventories to be built up in anticipation of future demand peaks.

While appearing deceptively simple, the SPCF model already involves a number of subtleties. The reader will have noticed that the output constraints (7.38) are written in inequality form; this is because writing them as equalities results in a non-convex feasible region (Merchant and Nemhauser 1978) as seen in the following example:

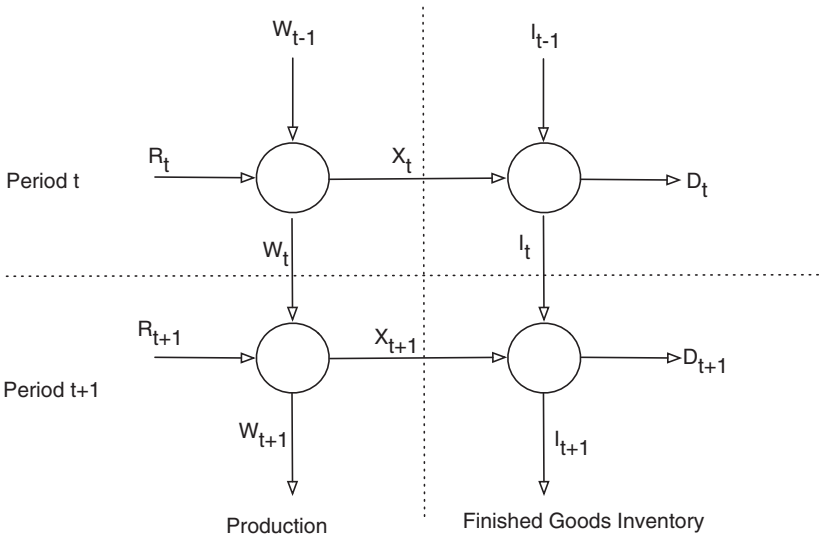

Fig. 7.3 Material Flows in Single-Product CF Model (Karmarkar 1989)

Example 7.1 Consider a two-period production planning problem with $D_1 = 3$, $D_2 = 9$, $I_0 = W_0 = 0$ and a clearing function

$$f(\Lambda_t) = \frac{10\Lambda_t}{10 + \Lambda_t} = \frac{10(W_{t-1} + R_t)}{10 + W_{t-1} + R_t}, \quad \Lambda_t \geq 0 \tag{7.40}$$

Consider the two solutions Y^1 and Y^2 summarized in Table 7.1.

Now consider a solution $Y^3 = 0.3Y^1 + 0.7Y^2$. The reader can easily verify that Y^3 satisfies the material balance constraints (7.36) and (7.37). However, $X_3^1 = 0.3X_1^1 + 0.7X_1^2 = 5.67 < f\left(0.3\left(R_1^1 + W_0\right) + 0.7\left(R_1^2 + W_0\right)\right) = f(15.5) = 6.07$. Similarly, $X_2^3 = 7.51 < f\left(0.3(90) + 0.7(40)\right) = 7.84$. Thus Y^3 is not a feasible solution when the clearing function constraints (7.38) are enforced at equality, indicating a non-convex feasible region.

The slack variables associated with constraints (7.38) represent a situation where the resource is not producing the maximum output it is capable of given the workload available to it; it is holding back some WIP that it is capable of converting into output because of adverse consequences in future periods. The following example illustrates this behavior of the SPCF model.

Example 7.2 Consider a two-period instance of the SPCF model with $r = 1$, $w = 2$, $h = 3$, $c = 1$, $D_1 = 9$, and $D_2 = 3$. Assume the same workload-based clearing function used in Example 7.1. The optimal solution to this instance is illustrated in Fig. 7.4.

Table 7.1 Data for Example 7.1

Solution	R_1^i	R_2^i	W_1^i	W_2^i	Λ_1^i	Λ_2^i	X_1^i	X_2^i	I_1^i	I_2^i
Y^1	5	65	1.67	57.97	5	66.67	3.33	8.70	0.33	0.03
Y^2	20	10	13.33	16.33	20	23.33	5.67	7.51	2.67	1.17

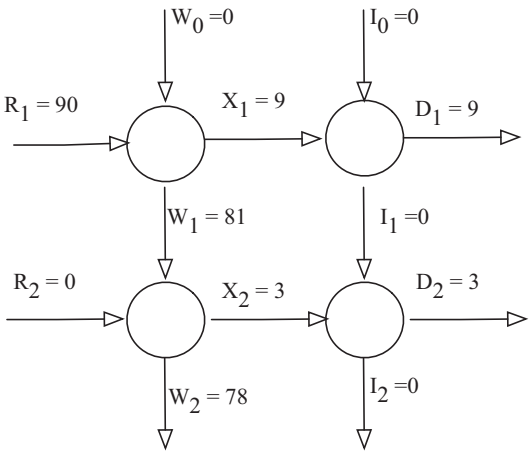

Fig. 7.4 Example of optimal solution with slack in CF constraints for Example 7.2

The high demand in period 1 requires the release of a large amount of work in that period to raise the workload to a level allowing output to meet the demand. The concave shape of the clearing function results in a large amount of WIP remaining at the end of period 1. However, the low demand in period 2 can be met with only three units of production. Since processing a unit of WIP to pass it into finished goods inventory incurs a total unit cost of $c = 1$ for production and $h = 3$ for holding the resulting finished inventory, it is cheaper to hold the excess material as WIP, resulting in a production of $X_2 = 3 < f(81) = 8.9$ units. Note that the behavior remains the same even if $c = 0$; simply having $h > w$ is sufficient.

Carey (1987) shows that constraints (7.38) will be satisfied as equalities as long as the marginal cost $c + (h - w)$ of moving material from WIP to finished goods in the absence of demand for it is nonpositive. To see this, note that in the network representation in Fig. 7.4, there are exactly two arcs incident out of the node corresponding to the WIP balance equation (7.36). Material in WIP at the start of period t is either retained in WIP in the next period (the vertical arc), or produced and moved to FGI (the horizontal arc). When a unit of WIP is converted into output and remains in finished inventory at the end of the period, the production cost of c is incurred, and the total holding cost in that period increases by $(h - w)$. An item for which there is external demand in the period will not incur the FGI holding cost h, and will be produced even if $c > w$ since otherwise demand will not be met, resulting in an infeasible solution in the absence of backlogging.

This holding back behavior can be explained in the context of traffic modeling as avoiding the release of traffic from one road segment to prevent congestion at downstream segments, say using traffic lights to regulate the flow of traffic. However, in production systems it is uncommon to hold WIP within the production process (as opposed to inventory points where intermediate products can be stored) without processing it if the capacity to process it is available, unless it is on hold due to quality or engineering problems. Thus this holding back behavior needs to be considered when implementing clearing function based planning models. The simplest approach is to set WIP holding costs sufficiently high ($w > c + h$ in this example) to ensure it is cheaper to move material downstream rather than retain it in the queue for a given process as WIP; after all, this is how production managers seem to behave in practice. However, this contradicts conventional cost accounting practice under which the holding cost of an item increases as it moves towards completion, due to the increasing value added during production. While it can be justified in some situations, such as semiconductor wafer fabrication where the high cost and limited availability of clean room space makes holding strategic inventory inside the factory undesirable, the manipulation of costs in this manner needs to be considered carefully in the context of the economics of the production system under study.

We have just seen that the SPCF model exhibits interesting behavior when restricted to a single-stage production system. We now explore its obvious extensions to multistage single-product and single-stage multiple product systems.

7.5 Multistage Single-Product Systems

The SPCF model (7.35)–(7.39) can be extended to multistage single-product environments in a straightforward manner by defining an index n denoting the stage of the production process. Thus a product is assumed to require a total of $n = 1, 2, ..., N$ operations whose sequence, or routing, is known and deterministic. However, this requires addressing the issue raised above of whether strategic inventory can be held between production stages or only at the output of the final stage. We shall first examine the model assuming such inventory cannot be held at intermediate locations, and then briefly discuss the case where such inventory can be held. For simplicity of exposition, we shall assume that each stage of the production process or routing corresponds to a distinct resource, each represented by its own clearing function. The extension to reentrant flows, where the product may undergo multiple operations at the same workcenter, is straightforward and can be addressed in exactly the same manner used for conventional models with fixed lead times (Leachman 2001; Kacar et al. 2016).

The parameters and decision variables remain the same as those in the SPCF model, except for the addition of an index n denoting the stage of the production process to which they refer. Demand can only be met with the output of stage N, and we shall again assume no backlogging of missed demand.

Our first model assumes that no inventory can be held within the production unit for tactical purposes such as anticipation of a future demand peak; such inventory is only held after stage N and consists of finished goods that can be used to meet demand. In this situation, work is released into the system at stage $n = 1$; the input Y_{nt} to stages $n > 1$ in a period t is given by the output of the previous stage in that period, i.e., $Y_{nt} = X_{n-1,t}$ in the notation of Chap. 6. This single-product multistage clearing function model (SPMCF) can be written as follows:

$$\min \sum_{t=1}^{T} \left[rR_t + hI_t + \sum_{n=1}^{N} \left(c^n X_t^n + w^n W_t^n \right) \right] \tag{7.41}$$

subject to

$$W_t^1 = W_{t-1}^1 + R_t - X_t^1, \quad t = 1,\ldots,T \tag{7.42}$$

$$W_t^n = W_{t-1}^n + X_t^{n-1} - X_t^n, \quad n = 2,\ldots,N, \quad t = 1,\ldots,T \tag{7.43}$$

$$I_t = I_{t-1} + X_t^N - D_t, \quad t = 1,\ldots,T \tag{7.44}$$

$$X_t^n \le f^n\left(\Lambda_t^n\right), \quad n = 1,\ldots,N, \quad t = 1,\ldots,T \tag{7.45}$$

$$I_t, W_t^n, X_t^n, R_t \ge 0, \quad n = 1,\ldots,N, \quad t = 1,\ldots,T \tag{7.46}$$

The decision variables and constraints in this model are analogous to those in the single-stage SPCF model (7.35)–(7.39). The WIP balance constraint (7.43) is written for stages 2, …, N, where the output from the previous stage $n-1$ provides the input of new work entering the stage. The WIP balance constraint (7.42) for Stage 1 is written using the release variables R_t representing external releases of new work into the production unit. Constraints (7.44) represent the material balance equations for the finished goods inventory held after the final stage N.

The objective function of this model is straightforward; our interest lies in the constraint set which attempts to model the behavior of a multistage production unit. The following example illustrates the behavior of these constraints.

Example 7.3 To illustrate the behavior of the constraint set (7.42)–(7.46), consider a serial production system consisting of five identical stages. Each stage is modeled by the workload-based clearing function $f(\Lambda_t) = 10\Lambda_t/(10 + \Lambda_t)$ used in the previous examples. Assuming $W_0^n = 0$ for all $n = 1, \ldots, N$, we release $R_1 = 10$ units of work into the first stage in the first period, with $R_t = 0$ for all remaining $t > 1$. Table 7.2 shows the evolution of the system state and output over time, while Fig. 7.5 illustrates the output of each stage.

Several interesting observations emerge from Table 7.2. 16.7% of the material released at the start of period 1 exits the overall system in the period in which it is released, traversing all five stages in a single period. This is analogous to Equation (4.6) in the discussion of load-oriented order release that estimates the fraction of the workload released in a certain period that traverses the first n workcenters on its routing within the same period. Given our assumption of instantaneous material transfer between stages and the fact that all stages are empty at the start of period 1, this behavior seems reasonable. It requires slightly more than six periods for all material released to exit the system. The small ending WIP levels at stages 4 and 5 and the end of period 6 are due to the fact that $f(\Lambda_t) < \Lambda_t$ for the CF in the example.

In order to estimate the average cycle time at each stage in each period, we shall use the expression

$$T^n = \begin{cases} 0.1\Lambda_t^n, \text{if } \Lambda_t^n < 0.25 \\ \dfrac{\left(W_{t-1}^n + W_t^n\right)}{2X_t^n}, \text{otherwise} \end{cases} \tag{7.47}$$

Assuming that all quantities are given in units of product, the clearing function implies that the maximum possible output from each stage in a planning period is 10 units, for an average processing time per unit of 0.1 periods. The first expression represents the situation where the workload is sufficiently low that the entire available workload Λ_t^n can be converted into output in the same period. This is a slight approximation, since the slope of the CF at the origin is equal to 1 and is decreasing in Λ_t^n; however, at $\Lambda_t^n = 0.25$ the clearing function posits an output of 0.2439 units of product, an error of 2.5%. The second term estimates cycle time using Little's Law, where the time-average WIP level in a period is estimated as the arithmetic

Table 7.2 Response of Empty System to Step Input

t	R_t	Λ_1	X_1	W_1	Λ_2	X_2	W_2	Λ_3	X_3	W_3	Λ_4	X_4	W_4	Λ_5	X_5	W_5
0				0			0			0			0			0
1	10	10.000	5.000	5.000	5.000	3.333	1.667	3.333	2.500	0.833	2.500	2.000	0.500	2.000	1.667	0.333
2	0	5.000	3.333	1.667	5.000	3.333	1.667	4.167	2.941	1.225	3.441	2.560	0.881	2.894	2.244	0.649
3	0	1.667	1.429	0.238	3.095	2.364	0.732	3.589	2.641	0.948	3.522	2.605	0.917	3.254	2.455	0.799
4	0	0.238	0.233	0.006	0.964	0.879	0.085	1.827	1.545	0.282	2.462	1.976	0.487	2.775	2.172	0.603
5	0	0.006	0.006	0.000	0.090	0.090	0.001	0.372	0.359	0.013	0.845	0.779	0.066	1.382	1.214	0.168
6	0	0.000	0.000	0.000	0.001	0.001	0.000	0.014	0.014	0.000	0.080	0.079	0.001	0.247	0.241	0.006
7	0	0.000	0.000	0.0000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.007	0.007	0.000
8	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

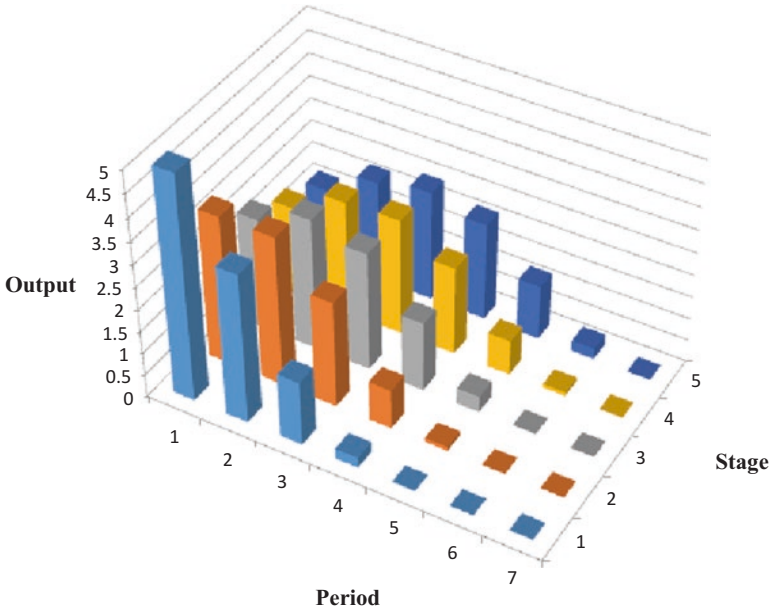

Fig. 7.5 Output by Stage and Period for Empty System in Example 7.3

average of its beginning and ending WIP levels. These estimates of cycle time are clearly crude averages; in particular, the use of Little's Law implies that the queue representing each stage in each period is in steady state, which requires, at the very least, long planning periods. If required, a relationship analogous to (6.5) can be used.

However, within these limitations, the results are still interesting, as shown in Fig. 7.6. The cycle time estimates at each stage increase in the early periods, as material arrives, and then decrease as the released material flows out of the system and is not replaced. The cycle time at each stage varies over time, highlighting the difficulties of using exogenous lead times in planning models. If we were to assume that each stage had a fixed lead time of 1 period and a maximum production capacity of 10 units per period—compatible with the clearing function used in the example—each stage n would produce an output of 10 units in period n, a completely different profile from that illustrated in Fig. 7.5.

For comparison, consider the situation illustrated in Table 7.3 and Fig. 7.7, where we again release 10 units into the system at the start of period 1, but each stage has 10 units in WIP at the start of that period. The combination of previous WIP and new releases results in a workload of $\Lambda_n = 20$ units at each stage n at the start of period 1. It now requires 12 periods for all work to exit the system. The output of all stages decreases over time, since the material released at the start of period 1 increases the output of all stages in that period, and hence moves material downstream to all stages in the subsequent periods. The additional 10 units of input at the

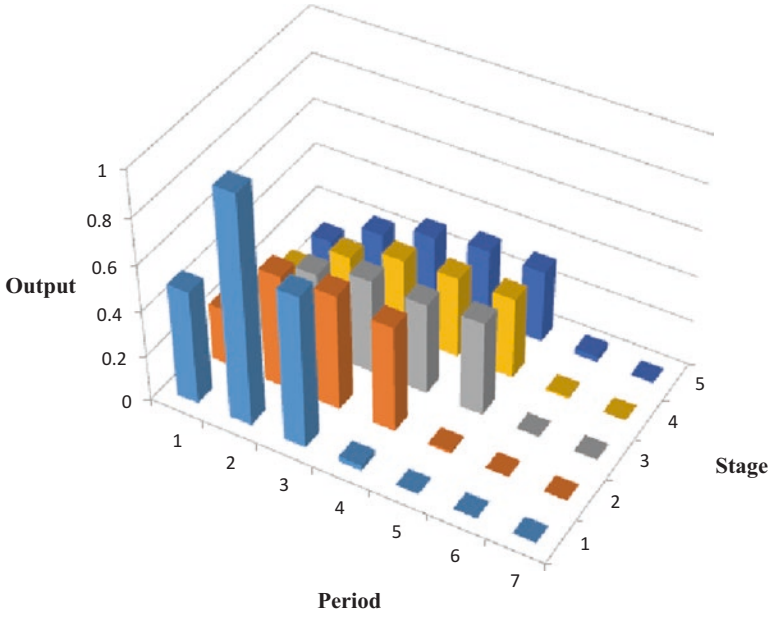

Fig. 7.6 Estimated Cycle Times for Example 7.3

start of period 1 increase the output of each subsequent stage in period 1 by less than 2 units; note that if there were no new releases, the output of each stage in period 1 would be 5 units. However, this does not imply that 16% of the newly released material completes all its processing in period 1. If we assume first-in-first-out processing at each stage, no new material is processed at Stage 1 in period 1; there are 10 units of WIP at the start of the period, of which only 6.667 units are converted into output. This is due to the relatively flat clearing function, which requires $\Lambda = 1000$ units to achieve an output of 9.9 units (Fig. 7.8).

The cycle times are now substantially higher than was the case with an empty system. The relatively slow decrease in the cycle times at all stages in periods 1 through 3 is noteworthy; after period 5, though, as the workload decreases the cycle time decreases rapidly. The relative stability of the cycle times in the early periods provides some insight into why fixed lead time models can work well under many situations: as long as the workload does not vary greatly from period to period, cycle times may remain stable, allowing a fixed lead time to provide a sufficiently accurate solution, especially if fractional lead times as suggested by Hackman and Leachman (1989) are used (Kacar et al. 2016).

This example provides a qualitative illustration of the behavior of the constraints (7.42)–(7.46) that represent the behavior of the production system using clearing functions, particularly the strong differences from the fixed lead time models in Chap. 5. We now extend this model to the case of multiple products competing for capacity at the resource, which proves to be treacherous territory.

Table 7.3 Response of Loaded System to Step Input

t	R_t	Λ_1	X_1	W_1	Λ_2	X_2	W_2	Λ_3	X_3	W_3	Λ_4	X_4	W_4	Λ_5	X_5	W_5
0				10			10			10			10			10
1	10	20.000	6.667	13.333	16.667	6.250	10.417	16.250	6.190	10.060	16.190	6.182	10.009	16.182	6.181	10.001
2	0	13.333	5.714	7.619	16.131	6.173	9.958	16.233	6.188	10.045	16.197	6.183	10.014	16.184	6.181	10.003
3	0	7.619	4.324	3.295	14.282	5.882	8.400	15.926	6.143	9.784	16.157	6.177	9.980	16.180	6.180	10.000
4	0	3.295	2.478	0.817	10.879	5.210	5.668	14.994	5.999	8.995	15.979	6.151	9.828	16.150	6.176	9.974
5	0	0.817	0.755	0.062	6.423	3.911	2.512	12.906	5.634	7.272	15.463	6.073	9.390	16.047	6.161	9.886
6	0	0.062	0.061	0.000	2.573	2.047	0.527	9.318	4.824	4.495	14.213	5.870	8.343	15.756	6.117	9.639
7	0	0.000	0.000	0.0000	0.527	0.501	0.026	4.995	3.331	1.664	11.675	5.386	6.288	15.025	6.004	9.021
8	0	0.000	0.000	0.000	0.026	0.026	0.000	1.690	1.446	0.244	7.734	4.361	3.373	13.382	5.723	7.659
9	0	0.000	0.000	0.000	0.000	0.000	0.000	0.244	0.239	0.006	3.612	2.653	0.958	10.313	5.077	5.236
10	0	0.000	0.000	0000	0.000	0000	0.000	0006	0.006	0.000	0.964	0.879	0.085	6.115	3.795	2.320
11	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.085	0.084	0.001	2.404	1.938	0.466
12	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.467	0.446	0.021
13	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.021	0.021	0.000
14	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

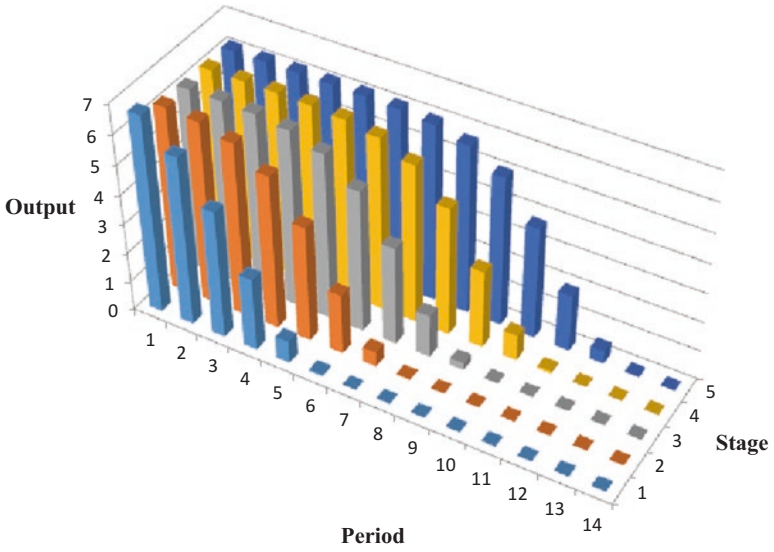

Fig. 7.7 Output of Loaded System in Example 7.3

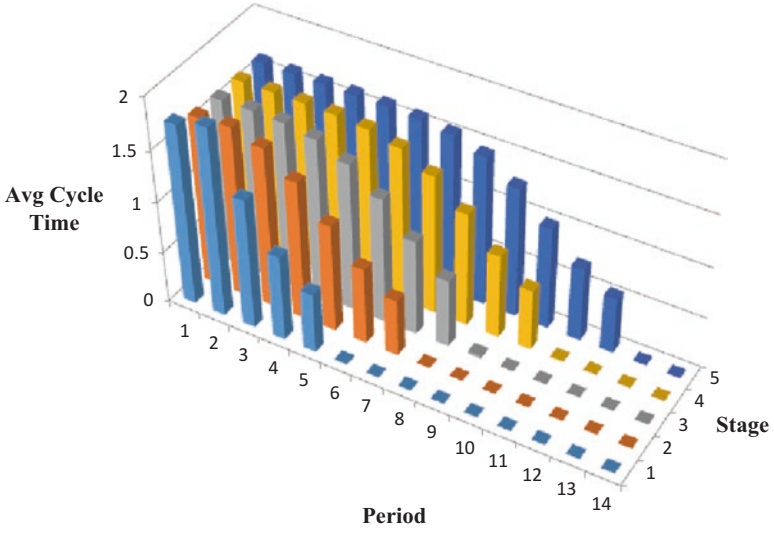

Fig. 7.8 Estimated Cycle Time of Loaded System Under Step Input

7.6 Single-Variable Clearing Functions with Multiple Products

The presence of multiple items brings the need to allocate the output of the resources among the different items. The use of a single-variable clearing function implies that the output of a resource is determined by its total workload, and hence that the

amount of output of each item must depend in some way on the workload of the other items. On the shop floor, the mix of items produced for a given workload of each item is determined by events on the shop floor, such as the arrival times of specific jobs at specific machines as well as scheduling and dispatching decisions. Since these operational policies are internal to the production unit, and hence not transparent to the planning level, it is difficult to model these directly in the planning problem. Even if the performance induced by these policies in the production unit could be described in computationally tractable models, it is not evident that it would be beneficial to do so, since local management will have the most current information on the status of the shop floor, and is responsible for managing the production unit in the face of changing local conditions. Hence a reasonable objective for the planning problem is to produce a release plan for each production unit that does not violate the basic constraints viewed by management as essential for the release plan to be usable.

One such set of constraints that has been widely discussed in the context of both production planning and traffic modeling is the maintenance of basic continuity conditions on the material flow. In both the traffic and production contexts, these can be expressed as a requirement that material entering the production unit earlier ought to exit earlier. Some deviation from this condition at the level of individual orders is clearly possible, and even desirable, in practice due to the ability of local management to expedite the processing of some jobs over others. Hence it ought to be sufficient for planning models to satisfy this requirement on average, while avoiding gross violations. Several sets of necessary and sufficient conditions for this first in first out (FIFO) property derived by Carey (1992) were discussed in Chap. 6, noting that they all lead to non-convex feasible regions.

We shall begin our discussion of multi-item models with single-variable clearing functions by presenting a naive extension of the SPCF model (7.35)–(7.39), to illustrate the difficulties that arise. We then discuss several solution approaches, most suggested in the context of traffic modeling (Carey 1992; Carey and Subrahmanian 2000a, b) which result in non-convex formulations. Finally, we present the allocated clearing function (ACF) model of Asmundsson et al. (2006, 2009), which provides a workable solution to these difficulties in the limited context of a single-variable clearing function.

7.6.1 Difficulties with Multiple Items

At first sight, extending the SPCF model to multiple items appears quite straightforward: we should add an item index i, write WIP balance and finished goods inventory balance equations for each item and add a clearing function constraint shared across all items. We use the following notation in addition to that already defined:

Indices:

i: item index, $i = 1,\ldots, I$

Parameters:

c_i: unit production cost of item i
h_i: unit finished goods inventory holding cost for item i
w_i: unit WIP holding cost for item i
r_i: unit cost of raw materials for item i, incurred upon release of the material to the resource
a_i: amount of time required on the resource to produce one unit of item i
$f(\Lambda_t)$: the clearing function, which we assume to be a concave, monotonically non-decreasing function of the total workload Λ_t
D_{it}: demand for item i in period t
I_{i0}: number of units of item i in finished goods inventory at start of period 1
W_{i0}: number of unprocessed units of item i in WIP at the start of period 1

Decision Variables:

X_{it}: output of item i in period t, in units of product
R_{it}: number of units of item i released to the resource in period t
I_{it}: number of units of item i remaining in finished goods inventory at the end of period t
W_{it}: number of units of item i in WIP at the end of period t
Λ_{it} : workload due to item i in period t, given by $a_i(R_{it}+W_{i,t-1})$ in units of time
Λ_t: total workload available to resource at the start of period t, given by

$$\Lambda_t = \sum_{i=1}^{I} a_i \Lambda_{it} \tag{7.48}$$

As in the previous examples, we assume time-stationary values of all parameters for simplicity of exposition. The Naive extension of the SPCF model to multiple items, which we shall refer to as the NSPCF model, can now be written as:

$$\min \sum_{t=1}^{T} \sum_{i=1}^{I} \left[r_i R_{it} + w_i W_{it} + c_i X_{it} + h_i I_{it} \right] \tag{7.49}$$

subject to:

$$I_{it} = I_{i,t-1} + X_{it} - D_{it}, \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.50}$$

$$W_{it} = W_{i,t-1} + R_{it} - X_{it}, \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.51}$$

$$\sum_{i=1}^{I} a_i X_{it} \le f\left(\Lambda_t\right), \quad t = 1,\ldots,T \tag{7.52}$$

$$R_{it}, X_{it}, I_{it}, W_{it} \ge 0, \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.53}$$

Table 7.4 Parameter values for Example 7.4

Item i	a_i	r_i	c_i	w_i	h_i	W_{i0}	I_{i0}
1	1	1	0	1	0	0	0
2	1	1	0	6	0	0	0

Table 7.5 Demand data for Example 7.4

Item i	Period 1	Period 2	Period 3	Period 4	Period 5
1	0	0	0	0	0
2	8	8	8	8	8

The decision variables and constraints of this model are completely analogous to those in the SPCF model. However, the treatment of the clearing function causes some interesting difficulties, which we illustrate in the following example.

Example 7.4 Consider a problem instance with two items, five periods, the parameter values shown in Table 7.4, and the demand data in Table 7.5. We again use the clearing function $f(\Lambda_t) = 10\Lambda_t/(10 + \Lambda_t)$ used in the previous examples.

Solving the model (7.49)–(7.53) yields the solution in Table 7.6. The problem is evident upon inspecting the results. There is no demand for item 1 in any period, and yet 29.93 units of Item 1 are released into the system, none of which is converted into output. The total workload generated by both products is used to meet the demand for item 2 with minimum WIP holding cost. Note that in periods 1 and 2, the model elects to produce less output than the clearing function allows: 8.18 units in period 1 as opposed to the 8.35 units the clearing function allows for the available workload of $\Lambda_t = 50.59$ units.

The problem is now clear: the WIP of item 1 is being held stationary in the system to artificially raise the available workload and permit the expensive item 2 to pass through the system rapidly. The planned releases of item 2 cannot, on their own, generate sufficient workload to produce the planned output of item 2. Simply parking idle WIP on the shop floor is increasing the output capacity of the system!

The reason for this behavior is also apparent. There is nothing in the model that links the output of the system to the composition of the workload enabling that output. This can be interpreted as a violation of the no-passing condition mentioned above—we are allowing the new releases of item 2 to constantly overtake the material of item 1 released in period 1. While in any practical production system some overtaking will arise naturally through the operation of shop-floor scheduling and dispatching systems, the idea that holding inert, idle WIP in the system increases the capability of the resources is clearly unrealistic.

As seen in Example 7.1, the non-convexity of clearing function models for single-item formulations enforcing the clearing function constraints as equalities was identified by Merchant and Nemhauser (1978). Carey (1987) demonstrated that implementing the clearing (exit) functions as inequalities results in a convex optimization problem for the single-item case, and discusses the issue of flow controls, where the clearing function constraints may hold as strict inequalities. He shows that the holding back behavior discussed in Example 7.2 will be avoided if the

Table 7.6 Optimal solution for Example 7.4

Period	R_1	R_2	W_1	W_2	Λ_1	Λ_2	Λ	$f(\Lambda)$	X_1	X_2	aX	I_1	I_2	D_1	D_2
0			0	0								0	0		
1	29.93	20.65	29.93	12.47	29.93	20.65	50.59	8.35	0.00	8.18	8.18	0.00	0.18	0	8
2	0.00	3.72	29.93	8.12	29.93	16.19	46.13	8.22	0.00	8.08	8.08	0.00	0.26	0	8
3	0.00	0.00	29.93	0.20	29.93	8.12	38.05	7.92	0.00	7.92	7.92	0.00	018	0	8
4	0.00	7.71	29.93	0.00	29.93	7.91	37.84	7.91	0.00	7.91	7.91	0.00	0.09	0	8
5	0.00	7.91	29.93	0.00	29.93	7.91	37.84	7.91	0.00	7.91	7.91	0.00	0.00	0	8

marginal cost of moving flow downstream (in our context, moving material from WIP to finished goods inventory) is lower than that of holding it at its current location. Carey discusses this issue in the context of modeling traffic flows and suggests a number of options. We digress briefly to discuss several of these, since they highlight a number of issues arising in optimization models involving flows through production networks. Our discussion follows that of Carey (1992), adapting the notation to the production planning models of interest in this work.

7.6.2 Enforcing Average No-Passing (FIFO) Behavior

Returning for a moment to the single-item problem, let R_{st} denote the amount of material released in period s that is converted into output in period t. Thus, in the notation of the SPCF model, we have

$$X_t = \sum_{s=1}^{t} R_{st} \tag{7.54}$$

One way of enforcing a no-passing condition is to ensure that material released earlier cannot be converted into output (i.e., transition from WIP to finished goods inventory) after material that is released later. This implies a condition that

$$R_{st} > 0 \rightarrow \sum_{\{s'<s,t'>t\}} R_{s',t'} = 0 \tag{7.55}$$

If we have multiple items $i = 1,...,I$, (7.55) must hold for each item i, as well as all pairs of items i and j. The explicit enforcement of this condition requires the use of integer variables to represent what are effectively disjunction constraints, resulting in computationally demanding integer programming models.

Intuition suggests that the no-passing property is likely to be violated if there are large changes in cycle times from one period to the next. This would suggest that as long as the cycle times (flow times in the traffic terminology) are "smooth enough" over time, violations of no-passing ought to be at least mitigated. As seen in Chap. 6, we can calculate the average cycle time for material released into the system in period s as

$$\bar{L}_s = \frac{\sum_{\tau=s}^{T} (\tau - s) R_{s\tau}}{\sum_{\tau=s}^{T} R_{s\tau}} \tag{7.56}$$

which is non-convex in the R_{st} variables (Carey 1992). Thus the average unit of work emerging as finished goods inventory at time $t + \bar{L}_t$ must have entered the system at time t. Carey (1992) then shows that the condition

$$\bar{L}_t \le \bar{L}_{t+1} + 1 \tag{7.57}$$

is necessary and sufficient to ensure no-passing of the average flows, and necessary but not sufficient to ensure no-passing on all individual work releases in the single-item case. This implies that the possibility of passing only arises when cycle times associated with the release periods are decreasing over time, i.e., the workload in the system is decreasing. As was the case for (7.55), in the presence of multiple items, this requires similar constraints to be written for each item i and all pairs of items i, requiring $O(I^2)$ additional non-convex constraints in each period where I is the number of items. In the presence of multiple items, we can enforce no-passing for all items by requiring that all pairs of items i and j have the same average cycle time, i.e.,

$$\bar{L}_{it} = \bar{L}_{jt} = \bar{L}_t, \quad \text{for all pairs of items} \quad i, j = 1,\ldots,I,\ i \ne j \tag{7.58}$$

where $\bar{L}_t$ denotes the average cycle time associated with material released in period t over all items $i = 1,\ldots,I$, and then enforcing (7.58).

A third approach to ensuring no-passing solutions is to assume that the production unit selects work for processing from the available workload without systematically prioritizing any item over any other. In this case, the mix of items converted into finished goods inventory in a period should match the mix of the items in the available workload, i.e., for all items i we should have

$$\frac{a_i X_{it}}{\sum_{i=1}^{I} a_i X_{it}} = \frac{\Lambda_{it}}{\sum_{i=1}^{I} \Lambda_{it}} \tag{7.59}$$

Although these conditions also result in non-convex constraints (Carey 1992), they form the basis for the allocated clearing function presented in the next section, which provides a computationally tractable approximate formulation that has proven effective for multi-item problems.

7.7 The Allocated Clearing Function (ACF) Model

The difficulties with the NSPCF formulation (7.49)–(7.53) arise because there is no constraint linking the output of each item in a period to the workload of that item available for processing in the period. This results in violation of the no-passing property, where workload of a cheap item is held immobile to allow rapid throughput of an expensive item without the need for high WIP levels of the latter. Clearly additional constraints of some sort are needed to address this situation, and we have discussed several possibilities in the previous section. However, all of these result in non-convex optimization models, which are computationally challenging to solve exactly for a proven global optimum. Hence some kind of approximation will be necessary.

To derive the ACF formulation, we consider the clearing function constraints (7.52) from the NSPCF model:

$$\sum_{i=1}^{I} a_i X_{it} \le f\left(\Lambda_t\right).$$

We wish to develop a set of constraints that relate the output X_{it} of each item i in period t to the workload Λ_{it} of that item in that period, while continuing to satisfy (7.52). To this end, we define a new set of variables Z_{it} as the fraction of total system output in period t allocated to item i in that period, i.e.,

$$Z_{it} = \frac{a_i X_{it}}{\sum_{j=1}^{I} a_j X_{jt}} \tag{7.60}$$

The definition of the Z_{it} implies that

$$\sum_{i=1}^{I} Z_{it} = 1 \tag{7.61}$$

The following constraint set is now equivalent to (7.52):

$$\begin{aligned} &a_i X_{it} \le Z_{it} f\left(\Lambda_t\right), \quad \text{forall} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \\ &\sum_{i=1}^{I} Z_{it} = 1, \quad t = 1,\ldots,T \end{aligned} \tag{7.62}$$

since summing the first set of constraints over all items i recovers (7.52). We can now incorporate the no-passing conditions (7.59) suggested by Carey (1992) to obtain the constraints

$$\begin{aligned} &a_i X_{it} \le Z_{it} f\left(\Lambda_t\right), \quad \text{forall} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \\ &\sum_{i=1}^{I} Z_{it} = 1, \quad t = 1,\ldots,T \end{aligned} \tag{7.63}$$

$$\frac{a_i X_{it}}{\sum_{i=1}^{I} a_i X_{it}} = \frac{\Lambda_{it}}{\sum_{i=1}^{I} \Lambda_{it}} = Z_{it}, \quad \text{forall} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.64}$$

The last constraint implies that

$$\Lambda_t = \frac{\Lambda_{it}}{Z_{it}} \tag{7.65}$$

yielding the constraint set

$$
\begin{aligned}
& a_i X_{it} \le Z_{it} f\left(\frac{\Lambda_{it}}{Z_{it}}\right), \quad \text{forall} \quad i = 1,\ldots,I \\
& \sum_{i=1}^{I} Z_{it} = 1 \\
& \frac{a_i X_{it}}{\sum_{i=1}^{I} a_i X_{it}} = \frac{\Lambda_{it}}{\sum_{i=1}^{I} \Lambda_{it}} = Z_{it}, \quad \text{forall} \quad i
\end{aligned}
\tag{7.66}
$$

The first constraint in (7.66) achieves our initial goal of obtaining a set of constraints that link the available workload Λ_{it} of each item in the period to the output of that item in the period. However, it looks like we now have some seriously non-convex constraints. The situation is not as bad as it appears at first sight, however. A standard result in convex analysis states that for any convex function *f(x),* its perspective *zf(x/z)* is also convex (Boyd and Vandenberghe (2009): 89). Hence the two constraints in (7.66) define a convex feasible region. However, we have seen in Chap. 6 that the last constraint results in a non-convex feasible region.

To develop an approximate constraint set that may be more tractable than (7.66), we relax the third constraint set, which leaves us with the constraints

$$
\begin{aligned}
& a_i X_{it} \le Z_{it} f\left(\frac{\Lambda_{it}}{Z_{it}}\right), \quad \text{forall} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \\
& \sum_{i=1}^{I} Z_{it} = 1, \quad t = 1,\ldots,T
\end{aligned}
\tag{7.67}
$$

The consequence of this relaxation is that the argument of the clearing function in the first constraints may not be accurate; we need not necessarily have $\Lambda_t = a_i \Lambda_{it}/Z_{it}$. This, in turn, introduces the possibility that the aggregate output constraint (7.52) may be violated if $a_i \Lambda_{it}/Z_{it} > \Lambda_t$ for some items *i*. To see that this is not the case, we need to show that the total output of all items *i* cannot exceed the aggregate output of the system implied by its total workload Λ_t, i.e.,

$$
\sum_{i=1}^{I} Z_{it} f\left(\frac{\Lambda_{it}}{Z_{it}}\right) \le f\left(\Lambda_t\right) \tag{7.68}
$$

We can write

$$
\sum_{i=1}^{I} Z_{it} f\left(\sum_{i=1}^{I} \frac{\Lambda_{it}}{Z_{it}}\right) \le f\left(\sum_{i=1}^{I} Z_{it}\left[\frac{\Lambda_{it}}{Z_{it}}\right]\right) = f\left(\sum_{i=1}^{I} \Lambda_{it}\right) = f\left(\Lambda_t\right) \tag{7.69}
$$

where the first inequality holds by the assumed concavity of *f(.)* and (7.61), the first equality from simple algebra and the second from the definitions of Λ_t and Λ_{it}. Since this assumes only the concavity of the clearing function *f(.),* it holds for any concave

clearing function. We can thus write the complete allocated clearing function formulation for a single-stage multi-item problem as follows:

$$\min \sum_{t=1}^{T}\sum_{i=1}^{I}\left[r_i R_{it} + c_i X_{it} + w_i W_{it} + h_i I_{it}\right] \tag{7.70}$$

subject to

$$W_{it} = W_{i,t-1} + R_{it} - X_{it}, \quad \text{for} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.71}$$

$$I_{it} = I_{i,t-1} + X_{it} - D_{it}, \quad \text{for} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.72}$$

$$a_i X_{it} \le Z_{it} f\left(\frac{\Lambda_{it}}{Z_{it}}\right), \quad \text{for} \quad i = 1,\ldots,I, \quad T = 1,\ldots,T \tag{7.73}$$

$$\sum_{I=1}^{I} Z_{it} = 1, \quad t = 1,\ldots,T \tag{7.74}$$

$$R_{it}, X_{it}, W_{it}, I_{it}, Z_{it} \ge 0, \quad \text{for} \quad i = 1,\ldots,I, \quad t = 1,\ldots,T \tag{7.75}$$

It is important to clarify the precise nature of the approximation being used here. The approximation arises from the fact that we estimate the total output of the system $f(\Lambda_t)$ by $f\left(\frac{\Lambda_{it}}{Z_{it}}\right)$ in the constraints (7.68). If we retain the no-passing constraints (7.64), the estimate of $f(\Lambda_t)$is exact; however, retaining these constraints results in a non-convex optimization model. By relaxing (7.64), we allow the mix of the output, defined by the ratios $a_i X_{it} / \sum_{i=1} a_i X_{it}$ to deviate from the mix of available workload, determined by the ratios Λ_{it}/Λ_t. Thus the ACF model may decide to process a larger fraction of the workload of one item *i* at the expense of another item *j*. However, there are limits to what is possible, as discussed in the next section. The primary insight is that despite their rather intimidating appearance, constraints (7.73) and (7.74) define a convex feasible region, resulting in a convex feasible region for the overall model (7.70)–(7.75) as long as the clearing function is concave.

7.7.1 ACF Model with Piecewise Linearized Clearing Function

Piecewise linearizing the clearing function as in (7.27), we can approximate the convex constraints (7.73) with the linear constraints

$$\begin{aligned} &a_i X_{it} \le Z_{it}\left(\alpha^q \frac{\Lambda_{it}}{Z_{it}} + \beta^q\right) = \alpha^q \Lambda_{it} + Z_{it}\beta_{it}^q, \\ &\text{for} \quad i = 1,\ldots,I, \; q = 1,\ldots,Q, \; t = 1,\ldots,T \end{aligned} \tag{7.76}$$

Now the ACF formulation has come into its own: the piecewise linear approximation of the single-variable clearing function has resulted in a set of linear constraints, yielding a linear program. However, this comes at the price of a substantial increase in the size of the model. The nonlinear model (7.70)–(7.75) has $5IT$ decision variables and $3IT + T$ constraints, of which the IT constraints (7.73) are nonlinear. The piecewise linearized formulation requires IQT linear constraints to approximate the nonlinear constraints (7.73). As a point of reference, a model using exogenous lead times would require $2IT$ decision variables representing releases and finished inventories and IT capacity and finished inventory balance constraints. As might be expected, the effort to model congestion more effectively increases the computational effort required to solve the models.

The following example illustrates the operation of the ACF formulation.

Example 7.5 Consider a problem with $T = 14$ time periods and two products whose data is given in Table 7.7 below:

The c_i, r_i, and initial WIP and inventory levels for both products have been set to zero for simplicity of exposition. The demand data over the planning horizon is given in Table 7.8, and the data for the linear segments approximating the workload-based clearing function in Table 7.9.

The solution of the ACF model is summarized in Table 7.10. The reader can verify that there is no slack in the clearing function constraints in any period. The high inventory holding costs require the model to operate with as little finished inventory as possible. In periods 1 through 3 and periods 11 through 14, only one product is produced. The shaded cells for periods 4 through 10 represent periods in which both products are in production. In periods 8 and 10 a higher fraction of the aggregate output, represented by the Z_{it} variables, is allocated to Product 1 than would be implied by the WIP fraction. For example, in period 10, Product 1 makes up 47% of the average WIP and yet is allocated 68% of the output capacity. This illustrates

Table 7.7 Parameters for Example 7.5

Product	a_i	c_i	w_i	h_i	r_i	W_{i0}	I_{i0}
1	2	0	6	5	0	0	0
2	4	0	11	10	0	0	0

Table 7.8 Demand Data for Example 7.5

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

Table 7.9 Clearing Function Parameters for Example 7.5

Segment c	Slope α^q	Intercept β^q
1	1	0
2	0.3	10
3	0	20

Table 7.10 Solution to ACF Model of Example 7.5

												Workload Fraction		Output Fraction				
Period	R_{1t}	R_{2t}	W_{1t}	W_{2t}	Λ_{1t}	Λ_{2t}	Λ_t	X_{1t}	X_{2t}	I_{1t}	I_{2t}	Prod. 1	Prod. 2	Z_{1t}	Z_{2t}	Ext. WL 1	Ext WL 2	Λ_t
0	0	0	0	0						0	0							
1	5.00	0.00	0.00	0.00	10.00	0.00	10.00	5.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	10.00	ZERO	10.00
2	5.00	0.00	0.00	0.00	10.00	0.00	10.00	5.00	0.00	0.00	0.00	1.00	0.00	0.70	0.30	14.29	0.00	10.00
3	5.00	0.00	0.00	0.00	10.00	0.00	10.00	5.00	0.00	0.00	0.00	1.00	0.00	0.70	0.30	14.29	0.00	10.00
4	5.00	0.86	0.00	0.00	10.00	3.43	13.43	5.00	0.86	0.00	0.86	0.74	0.26	0.70	0.30	14.29	11.43	13.43
5	5.00	1.07	0.00	0.00	10.00	4.29	14.29	5.00	1.07	0.00	1.93	0.70	0.30	0.70	0.30	14.29	14.29	14.29
6	3.00	2.07	0.00	0.00	6.00	8.29	14.29	3.00	2.07	0.00	0.00	0.42	0.58	0.42	0.58	14.29	14.29	14.29
7	6.67	5.00	2.67	2.00	13.33	20.00	33.33	4.00	3.00	0.00	0.00	0.40	0.60	0.40	0.60	33.33	33.33	33.33
8	7.10	1.33	3.84	1.33	19.54	13.33	32.87	5.93	2.00	0.93	0.00	0.59	0.41	0.60	0.40	32.57	33.33	32.87
9	8.23	0.97	4.83	0.92	24.14	9.20	33.33	7.24	1.38	2.17	0.38	0.72	0.28	0.72	0.28	33.33	33.33	33.33
10	0.00	1.78	0.00	1.08	9.66	10.80	20.46	4.83	1.62	0.00	0.00	0.47	0.53	0.68	0.32	14.29	33.33	20.46
11	0.00	2.35	0.00	0.00	0.00	13.71	13.71	0.00	3.43	0.00	0.43	0.00	1.00	0.00	1.00	ZERO	13.71	13.71
12	0.00	3.57	0.00	0.00	0.00	14.29	14.29	0.00	3.57	0.00	0.00	0.00	1.00	0.00	1.00	ZERO	14.29	14.29
13	0.00	3.00	0.00	0.00	0.00	12.00	12.00	0.00	3.00	0.00	0.00	0.00	1.00	0.00	1.00	ZERO	12.00	12.00
14	0.00	2.00	0.00	0.00	0.00	8.00	8.00	0.00	2.00	0.00	0.00	0.00	1.00	0.00	1.00	ZERO	8.00	8.00

the impact of relaxing the FIFO constraints (7.64) on the solution of the ACF model: Product 1 is being prioritized over Product 2 due to its higher WIP holding cost. This is possible because of the structure of the linearized clearing function constraints.

It is of particular interest to compare the output fraction of each product, given by the values of the Z_{it} variables in each period, with the workload fraction. The next to last two columns of the table show the extrapolated workload (Ext. WL) for each product, given by Λ_{it}/Z_{it} which can be compared to the actual workload Λ_t shown in the last column. In period 10, the extrapolated workload of Product 2 exceeds the actual workload by a considerable margin, while that of Product 1 falls below it. However, the weighted average of the two extrapolated workloads remains equal to the actual workload. In periods where the output fraction matches the workload fraction, extrapolated workload is equal to actual workload for both products. Thus the ACF model is increasing the output fraction of Product 2 by applying a smaller output fraction to larger extrapolated workload.

To see how this is accomplished, note that the output $a_i X_{it}$ of product i in period t can be decomposed into two components: one that is proportional to its workload WIP, given by $\alpha^q \Lambda_{it}$, and a portion of the intercept given by $Z_{it}\beta^q$. Unless $Z_{it} = 0$, the first component will always be produced in proportion to the available average WIP and the slope of the clearing function segment. However, the ACF model may distribute the β^c units of output due to the intercept of the linear segment in the manner yielding the best objective function value. The amount of this discretionary output, which can be allocated among products subject only to the WIP balance constraints, increases as the resource is more heavily utilized, leading to higher workload Λ_t, whichever way the clearing function has been formulated. However, since the Z_{it} variables must sum to 1, the total output in units of time allocated among the different products cannot exceed the disposable output β^q. Note that if the workload were

sufficiently high that only segment Q of the CF, with slope $\alpha^Q = 0$, were binding, the model could allocate output arbitrarily among products subject only to the WIP balance constraints, essentially replicating the N-SPCF model. However, it is easy to see that such a solution can never be optimal, as the same output can be achieved with lower total workload, and hence lower WIP holding costs.

In summary, the ACF formulation avoids the issues encountered with the N-SPCF formulation discussed in Sect. 7.6.1 by relaxing the constraint that output mix must exactly match the workload mix in each period. This allows flexibility in allocating output among the different products, but ensures positive production of all products with positive workload, avoiding the creation of capacity for one product by simply holding static WIP of another. It is by no means a fully satisfactory solution, but it has been extensively tested over more than a decade since its first introduction, and has consistently produced satisfactory, consistent solutions that have in many cases outperformed the fixed lead time models described in Chap. 5. Recent results (Gopalswamy and Uzsoy 2018) have shown that as long as the CF used is concave, the ACF model can be extended to a second-order conic programming formulation which preserves the structure of the dual solution described in the following section, and also significantly reduces the variability of releases across time periods frequently observed with linear programming models.

7.7.2 Dual Solution of the ACF Model

Recall from Chap. 5 that any resource with utilization below 1 in any period will have slack in its capacity constraint for that period, resulting in a zero value for the associated dual variable. We now develop and analyze the dual solution for the multistage clearing function model equivalent to that analyzed in Sect. 5.4. The analysis in this section is based on that in Kefeli and Uzsoy (2016), modified slightly for consistency with the discussion in Chap. 5. We shall consider the production system consisting of K resources in series modeled in (5.42)–(5.55) for the case of fixed lead times. No strategic inventory of intermediate products is held between stages inside the production unit; the output of all stages $k = 1, ..., K-1$ except the final one moves directly into the WIP of the next stage $k+1$. Raw material is released into stage 1, and material completing processing at stage K enters finished goods inventory from where it can be withdrawn to meet demand. We represent each stage with its own workload-based clearing function $f_k(\Lambda_{kt})$, where Λ_{kt} denotes the planned workload at stage k in period t. To implement the ACF model $f_k(\Lambda_{kt})$ will be approximated using the piecewise linearization (7.27) as

$$f_k\left(\Lambda_{kt}\right) = \alpha_k^q \Lambda_{kt} + \beta_k^q, \quad q = 1,\ldots,Q \tag{7.77}$$

To facilitate the sometimes extensive notation, we shall denote the set of all products by I and the index set of all linear segments approximating the clearing function for resource k as Q, assuming without loss of generality that all resources are

approximated by the same number of linear segments. Using this notation, we write the ACF formulation as follows:

$$\min \sum_{t=1}^{T} \sum_{i \in I} \left(h_{it} I_{it} + r_{it} R_{it} + \sum_{k=1}^{K} \left(p_{it}^{k} X_{it}^{k} + w_{it}^{k} W_{it}^{k} \right) \right) \tag{7.78}$$

subject to

$$W_{it}^{1} = W_{i,t-1}^{1} + R_{it} - X_{it}^{1}, \quad i \in I, \, t = 1,\ldots,T \tag{7.79}$$

$$W_{it}^{k} = W_{i,t-1}^{k} + X_{it}^{k-1} - X_{it}^{k}, \quad i \in I, \quad k = 2,\ldots,K, \quad t = 1,\ldots,T \tag{7.80}$$

$$I_{it} = I_{i,t-1} + X_{it}^{K} - D_{it}, \quad i \in I, \quad t = 1,\ldots,T \tag{7.81}$$

$$a_i^1 X_{it}^{1} \le \alpha_q^1 a_i^1 \left(R_{jt} + W_{i0}^{1} \right) + Z_{it}^{1} \beta_q^1, \quad i \in I, \quad q \in Q, \quad t = 1,\ldots,T \tag{7.82}$$

$$a_i^k X_{it}^{k} \le \alpha_q^k a_i^k \left(X_{it}^{k-1} + W_{i,t-1}^{k} \right) + Z_{it}^{k} \beta_q^k, \quad k = 2,\ldots,K, \quad i \in I, \quad q \in Q, \quad t = 1,\ldots,T \tag{7.83}$$

$$\sum_{i \in I} Z_{it}^{k} = 1, \quad k = 1,\ldots,K, \quad t = 1,\ldots,T \tag{7.84}$$

$$X_{it}^{k}, R_{it}, I_{it}, W_{jt}^{k}, Z_{it}^{k} \ge 0, \quad i \in I, \quad k = 1,\ldots,K, \quad t = 1,\ldots,T \tag{7.85}$$

Rewriting this in the cumulative form to eliminate the I_{it} and W_{it}^{k} variables and dropping constants from the objective function, we obtain

$$\min \sum_{t=1}^{T} \sum_{i \in I} \left[\begin{array}{l} \left(\left(r_{it} + \sum_{\tau=t}^{T} w_{i\tau}^{1} \right) R_{jt} + \sum_{k=1}^{K-1} \left(p_{it}^{k} + \sum_{\tau=t}^{T} \left(w_{i\tau}^{k+1} - w_{i\tau}^{k} \right) \right) X_{it}^{k} \right. \\ \left. + \left(p_{it}^{K} + \sum_{\tau=t}^{T} \left(h_{i\tau} - w_{i\tau}^{J(i)} \right) \right) X_{it}^{K} \right) \end{array} \right] \tag{7.86}$$

subject to

$$\sum_{\tau=1}^{t} R_{i\tau} - \sum_{\tau=1}^{t} X_{i\tau}^{1} \ge -W_{i0}^{1}, \quad i \in I, \quad t = 1,\ldots,T \tag{7.87}$$

$$\sum_{\tau=1}^{t} X_{i\tau}^{k-1} - \sum_{\tau=1}^{t} X_{i\tau}^{k} \ge -W_{i0}^{k}, \quad i \in I, \quad k = 2,\ldots,K, \quad t = 1,\ldots,T \tag{7.88}$$

$$\sum_{\tau=1}^{t} X_{it}^{K} - \sum_{\tau=1}^{t} D_{it} \ge -I_{i0}, \quad i \in I, \quad t = 1,\ldots,T \tag{7.89}$$

$$a_i^1 X_{it}^{1} \le a_i^1 \alpha_q^1 \left(R_{it} + W_{i,t-1}^{1} \right) + Z_{it}^{1} \beta_q^1, \quad q \in Q, \quad i \in I, \quad t = 1,\ldots,T \tag{7.90}$$

$$a_i^k X_{it}^k \le a_i^k \alpha_q^k \left(X_{it}^{k-1} + W_{i,t-1}^k \right) + Z_{it}^k \beta_q^k, \quad k = 2,\ldots,K, \quad q \in Q, \quad i \in I, \quad t = 1,\ldots,T \tag{7.91}$$

$$\sum_{i=1}^{I} Z_{it}^k = 1, \quad k = 1,\ldots,K, \quad t = 1,\ldots,T \tag{7.92}$$

$$R_{it}, X_{it}^k \ge 0, \quad k = 1,\ldots,K, \quad i \in I, \quad t = 1,\ldots,T \tag{7.93}$$

Analysis of the ACF model is simplified by defining two dummy resources 0 and $K + 1$, where resource $K + 1$ represents the arrival of the material in the finished goods inventory. Resource 0, on the other hand, represents the release of the raw material of product i into the line. Thus we define $p_{it}^0 = r_{it}$ for all products $i \in I$, and $w_{it}^0 = 0$. Similarly $w_{it}^{K+1} = h_{it}$ for all $i \in I$ and $t = 1,\ldots,T$, implying that $X_{it}^0 = R_{it}$ in the current notation. The formulation can now be written as follows:

$$\min \sum_{t=1}^{T} \sum_{i \in I} \sum_{k=0}^{K} \left(p_{it}^k + \sum_{\tau=t}^{T} \left(w_{i\tau}^{k+1} - w_{i\tau}^k \right) \right) X_{it}^k \tag{7.94}$$

subject to

$$\sum_{\tau=1}^{t} X_{i\tau}^K \ge \sum_{\tau=1}^{t} D_{i\tau} - I_{i0}, \quad i \in I, \quad t = 1,\ldots,T \qquad \left(\Gamma_{it}^{K+1} \right) \tag{7.95}$$

$$\sum_{\tau=1}^{t} X_{i\tau}^{k-1} - \sum_{\tau=1}^{t} X_{i\tau}^k \ge -W_{i0}^k, \quad i \in I, \quad t = 1,\ldots,T, \quad k = 1,\ldots,K \qquad \left(\Gamma_{it}^k \right) \tag{7.96}$$

$$-\alpha_q^k a_i^k \sum_{\tau=1}^{t-1} X_{i\tau}^k - a_i^k X_{it}^k + \alpha_q^k a_i^k \sum_{\tau=1}^{t} X_{i\tau}^{k-1} + Z_{it}^k \beta_q^k \ge -\alpha_q^k a_i^k W_{i0}^k, \qquad \left(\sigma_{itq}^k \right) \tag{7.97}$$
$$i \in I, \quad k = 1,\ldots,K+1, \quad t = 1,\ldots,T, \quad q = 1,\ldots,Q$$

$$\sum_{i \in I} Z_{it}^k = 1, \quad k = 1,\ldots,K, \quad i \in I, \quad t = 1,\ldots,T \qquad \left(\lambda_t^k \right) \tag{7.98}$$

$$X_{it}^k \ge 0, \quad i \in I; \quad k = 0,\ldots,K+1, \quad t = 1,\ldots,T \tag{7.99}$$

$$Z_{it}^k \ge 0, \quad i \in I; \quad k = 1,\ldots,K, \quad t = 1,\ldots,T \tag{7.100}$$

with the Greek letters in parentheses denoting the dual variables associated with each constraint. The dual of the formulation (7.94)–(7.100) is given by:

$$\max \sum_{t=1}^{T} \left\{ \sum_{i \in I} \left[\left(\sum_{\tau=1}^{t} D_{i\tau} - I_{i0} \right) \Gamma_{it}^{K+1} - \sum_{k=1}^{K} W_{i0}^k \Gamma_{it}^k - \sum_{k=1}^{K} \sum_{q} \alpha_q^k a_i^k W_{i0}^k \sigma_{itq}^k \right] + \sum_{k=1}^{K} \lambda_t^k \right\} \tag{7.101}$$

subject to

$$\sum_{\tau=t}^{T}\left(\Gamma_{i\tau}^{k+1}-\Gamma_{i\tau}^{k}\right)-\sum_{q\in Q}a_i^k\sigma_{itq}^k-\sum_{\tau=t+1}^{T}\sum_{q\in Q}\alpha_q^k a_i^k\sigma_{i\tau q}^k+\sum_{\tau=t}^{T}\sum_{q\in Q}\alpha_q^{k+1}a_i^{k+1}\sigma_{i\tau q}^{k+1}$$
$$\le p_{it}^k+\sum_{\tau=t}^{T}\left(w_{i\tau}^{k+1}-w_{i\tau}^k\right), i\in I,\ t=1,\ldots,T-1,\ k=0,\ldots,K\quad \left(X_{it}^k\right) \tag{7.102}$$

$$\Gamma_{iT}^{k+1}-\Gamma_{iT}^k-\sum_{q\in Q(k)}a_i^k\sigma_{iTq}^k+\sum_{q\in Q(k+1)}\alpha_q^{k+1}a_i^{k+1}\sigma_{iTq}^{k+1}\le p_{iT}^k+w_{iT}^{k+1}-w_{iT}^k,$$
$$i\in I,\quad k\in K\quad \left(X_{iT}^k\right) \tag{7.103}$$

$$\lambda_t^k+\sum_{q\in Q}\beta_q^k\sigma_{itq}^k\le 0,\quad i\in I,\quad k=1,\ldots,K,\quad t=1,\ldots,T\quad \left(Z_{it}^k\right) \tag{7.104}$$

$$\begin{aligned}&\Gamma_{it}^k\ge 0\quad i\in I,\quad k=1,\ldots,K+1\\&\sigma_{itq}^k\ge 0\quad i\in I,\quad k=1,\ldots,K\\&t=1,\ldots,T,\quad q\in Q\end{aligned} \tag{7.105}$$

$$\lambda_t^k\quad \text{free insign},\quad t=1,\ldots,T,\quad k=1,\ldots,K \tag{7.106}$$

with the associated primal variable indicated in parentheses next to each dual constraint. In the FLT model, the dual price of capacity is directly accessible as the dual variable $\hat{\sigma}_{kt}$ associated with the capacity constraints. Hence it is meaningful to refer to $\hat{\sigma}_{kt}$ as the dual price of capacity at workcenter k. The situation for the ACF model is more complex. The CF does not represent the "capacity" of the system in the sense of an upper limit on output; rather, it represents the relationship between expected workload and expected output at each workcenter k. Constraints (7.96) ensure that WIP is nonnegative in all periods, while (7.97) ensure that the output in each period is consistent with the capabilities of the workcenter described by its CF. Thus the dual variables Γ_{it}^k associated with (7.96) will only be nonzero when these constraints are tight at optimality, i.e., when workcenter k has no WIP of product i on hand at the end of period t. This is achieved when the cumulative output of product i at resource k in period t and the cumulative input of that item to that workcenter differ by W_{i0}^k, the initial WIP of product i at resource k at the start of the planning horizon. Thus the Γ_{it}^k can be interpreted as the cost impact in period t of a unit change in the initial WIP level W_{i0}^k. As implied by the dual objective (7.101), if all initial WIP and FGI values are set to $I_{i0}=W_{i0}^k=0$, the Γ_{it}^k variables have no impact on the optimal solution value except via the artificial workcenter $K+1$ representing the finished inventory. The dual variables Γ_{it}^{K+1} represent the maximum amount the firm should be willing to pay for an additional unit of finished inventory of product i in period t, or, equivalently, the minimum price it should charge an additional unit of demand.

The right hand side of the primal constraints (7.97) that limit output of each product by the CF computes the total output, in units of time, of product i for a given workload level. Thus the dual variables σ_{itq}^k indicate the amount the firm should be

willing to pay for an additional time unit of output of product i from resource k in period t. Examining the dual constraint (7.102), recall that a_i^k is the time required to process one additional unit of product i on resource k, and let $q \in Q$ denote the linear segment of resource k's CF with slope α_q^k that is binding at optimality. Rearranging (7.102) as follows provides some insight:

$$\begin{aligned} &\sum_{\tau=t}^{T}\left(\Gamma_{i\tau}^{k+1}-\Gamma_{i\tau}^{k}\right) \\ &+\underbrace{\sum_{q\in Q}\alpha_q^k a_i^k \sigma_{itq}^k-\sum_{q\in Q} a_i^k \sigma_{itq}^k}_{\text{net impact in period t}} \\ &+\underbrace{\sum_{\tau=t}^{T}\sum_{q\in Q}\left(\alpha_q^{k+1} a_i^{k+1}\sigma_{i\tau q}^{k+1}-\alpha_q^k a_i^k \sigma_{i\tau q}^k\right)}_{\text{net impact for remainder of planning horizon}} \le p_{it}^k+\sum_{\tau=t}^{T}\left(w_{i\tau}^{k+1}-w_{i\tau}^k\right) \end{aligned} \tag{7.107}$$

The right hand side indicates that an additional unit of output of product i at resource k in period t will reduce the WIP at this resource by one unit while increasing the WIP at the next resource $k+1$ along product i's routing; it will also save the incremental production cost p_{it}^k. The left hand side represents the total reduction in the objective function value due to this allocation over the remainder of the planning horizon. The impact in the current period t is the value of the additional output that can be generated from resource $k+1$, net of the value of the output from resource k, and the impact in the remainder of the planning horizon in a similar fashion. Thus the price paid by the firm for the additional output allocation should not exceed the cost savings from the purchase of the additional allocation.

In an optimal solution to the formulation (7.94)–(7.100), in any period t the workcenters can be classified into three groups: congested workcenters where $\sum_i W_{it}^k > 0$, non-congested workcenters where $\sum_i W_{it}^k = 0$ and $\sum_i X_{it}^k > 0$, and idle workcenters where $\sum_i X_{it}^k = \sum_i W_{it}^k = 0$. We shall define congested, non-congested, and idle periods for a workcenter analogously, depending on which of the three states defined above (congested, non-congested, or idle) the workcenter is in during the period in question. Recall we assume all products i are processed on all workcenters $k \in K$.

During idle periods, there is no external release of any product into the workcenter k and no production at the preceding operation in the product's routing, i.e., $X_{jt}^{k-1} = 0$. Hence there is no production or WIP present for that product at that resource k. In non-congested periods, production takes place but no WIP is carried from one period to the other. In this case, the workcenter is operating at sufficiently low utilization that all material arriving from previous operations or external releases is processed in the same period; the segment $q = 1$ with $\alpha_1^q = 1$ and $\beta_1^q = 0$ is tight at optimality. If a resource k is congested in some period t, on the other hand, the entire workload available to it in that period cannot be processed into output within the period, forcing some to be carried over to the next period as WIP. This means the system is operating at higher utilization and at least one segment of the CF with

index $q > 1$ is tight. Our analysis will focus on congested resources since these are where the differences with the FLT model are most clearly visible.

We define a congested interval $\Psi(k)$ for resource k to be a collection of consecutive congested periods starting with a period s and ending with a period $s' > s$, i.e., $\Psi(k) = \{s, s+1,\ldots,s'\}$ such that $\sum_i W^k_{i,s-1} = 0$ and $\sum_i W^k_{it} > 0$ for all $t \in \Psi(k)$ and $\sum_i W^k_{i,s'+1} = 0$.

Before we can apply complementary slackness to (7.101)–(7.106), we need several assumptions regarding the congested interval $\Psi(k)$. The complementary slackness conditions imply that $\Gamma^k_{it} = 0$ for all $t \in \Psi(k)$ since $W^k_{it} > 0$. We also assume that $W^{k+1}_{it} > 0$, so that we have $\Gamma^{k+1}_{it} = 0$, implying that the workcenter performing the next operation in the routing is also congested.

In order to be able to apply the complementary slackness conditions directly without the need to examine a wide range of cases, we will restrict our attention to periods where the system is in regular operation, i.e., $R_{it} > 0$ and $X^{k-1}_{it} > 0$ for some product i. These assumptions imply that $X^k_{it} > 0 \forall t \in \Psi(k)$, i.e., if a product is present at a workcenter due to either external releases or output from preceding workcenters, there must be production of that product on the workcenter. Otherwise we can release the work in a later period and save the WIP holding cost. For brevity of exposition, we shall assume that the last period $T \notin \Psi(k)$; in this case constraints (7.103) become active and are subject to a similar analysis.

We now apply complementary slackness to (7.101)–(7.106) during a congested interval $\Psi(k)$. Under the assumptions just stated, (7.102) and hence (7.107) are tight at optimality for all $t \in \Psi(k)$, yielding

$$\sum_{\tau=t}^{T}\left(\Gamma^{k+1}_{i\tau} - \Gamma^k_{i\tau}\right) + \sum_{q\in Q}\alpha^k_q a^k_i \sigma^k_{itq} - \sum_{q\in Q} a^k_i \sigma^k_{itq} + \sum_{\tau=t+1}^{T}\sum_{q\in Q}\left(\alpha^{k+1}_q a^{k+1}_i \sigma^{k+1}_{i\tau q} - \alpha^k_q a^k_i \sigma^k_{i\tau q}\right)$$
$$= p^k_{it} + \sum_{\tau=t}^{T}\left(w^{k+1}_{i\tau} - w^k_{i\tau}\right), \quad t \in \Psi(k) \tag{7.108}$$

Equations (7.108) collectively define the dual behavior of the optimal σ^k_{itq} in a congested interval. It is immediately evident that, unlike the FLT model, the dual price σ^k_{itq} associated with output of any product i at resource k is related to that associated with the preceding workcenter $k-1$ in its routing. We now rearrange (7.108) in such a fashion that their meaning is clearer by defining the quantity

$$\Phi^k_{it} = a^k_i \sum_{q\in Q} \alpha^k_q \sigma^k_{itq} \quad \forall i \in I, \quad t \in \Psi(k) \tag{7.109}$$

and rewriting (7.108) as follows:

$$\sum_{\tau=s'+1}^{T}\left(\Gamma^{k+1}_{i\tau} - \Gamma^k_{i\tau}\right) + \sum_{q\in Q}\alpha^k_q a^k_i \sigma^k_{itq} - \sum_{q\in Q(k)}^{T} a^k_i \sigma^k_{itq} - \sum_{\tau=t}^{T}\Phi^k_{i\tau} + \sum_{\tau=t}^{T}\Phi^{k+1}_{i\tau}$$
$$= p^k_{it} + \sum_{\tau=t}^{T}\left(w^{k+1}_{i\tau} - w^k_{i\tau}\right) \quad i \in I,\ t \in \Psi(k) \tag{7.110}$$

Writing (7.110) for periods t and $t+1$ and subtracting yields

$$a_i^k\left(\sum_{q\in Q}\sigma_{i,t+1,q}^k-\sum_{q\in Q}\sigma_{itq}^k\right)-\left(\Phi_{i,t+1}^k-\Phi_{i,t}^k\right)=\left(w_i^{k+1}-w_i^k\right)-\left(\Phi_{it}^{k+1}-\Phi_{it}^k\right) \quad (7.111)$$

illustrating the fact that the dual price associated with additional output of product i at workcenter k in period t impacts the dual prices at the downstream workcenter $k+1$ in its routing as suggested by queueing theory (Hopp and Spearman (2008), Chap. 8). Note also that the right hand side of this expression represents the impact of moving a unit of output from resource k to resource $k+1$ in period t, while the left hand side reflects its impact across time, from period t to $t+1$.

For the first workcenter $k = 1$ in the common routing (7.110) implies that

$$\sum_{\tau=s'+1}^{T}\Gamma_{i\tau}^1+\sum_{\tau=t}^{T}\Phi_{i\tau}^1=p_{it}^0+\sum_{\tau=t}^{T}w_{it}^1 \quad i\in I,\ t\in\Psi\left(k\right) \quad (7.112)$$

Writing (7.112) for periods from s' back to s and solving recursively yields

$$\Phi_{it}^1=\left(p_{it}^0-p_{i,t+1}^0\right)+w_{it}^1 \quad (7.113)$$

Under time-stationary costs $\left(p_{it}^0=r_{it}=r_i,w_{it}^k=w_i^k,h_{it}=h_i=w_{K+1},p_{it}^k=p_i^k\right)$ this expression simplifies to $\Phi_{it}^1=w_i^1,i\in I,t\in\Psi\left(k\right)\setminus\left\{s'\right\}$.

The primal constraints (7.98) represent the fact that the expected total output a workcenter k can produce in a given period t with a specified workload Λ_t^k is bounded above by the value $f_k\left(\Lambda_t^k\right)$ of the CF. Therefore the dual variables λ_t^k associated with (7.98) represent the change in objective function obtained by changing the value of this upper limit, i.e., changing the expected output of the workcenter in a period for a given workload Λ_t^k. This can be interpreted as the impact on the objective function value of having one additional time unit of output available in period t for allocation among the different products i, thus increasing the disposable output β_q^k (again in units of time) available for allocation by the Z_{it} variables. Although the dual variables λ_t^k are free in sign as a result of constraint (7.98) being defined as an equality, in any optimal solution these variables will only take negative values since an increase in the right hand side of (7.98) cannot yield an increase in the objective function value. Applying complementary slackness to (7.98), we get:

$$\lambda_t^k=-\sum_{c\in Q}\beta_c^k\sigma_{itc}^k, \quad i\in I,\ k\in K,\ t=1,\ldots,T \quad (7.114)$$

Thus at optimality output at each resource k is allocated among products to equalize the marginal value of the capacity allocated to each, in a manner consistent with the marginal value of additional output of each product i in each period t, given by the σ_{iqt}^k. Hence in our numerical experiments below, we shall use this quantity λ_t^k as the analog of the dual price of capacity derived for the FLT model.

Example 7.6 To see the difference in the behavior of the dual prices related to capacity, we compare the dual solution of the two-product single-stage problem in Example 7.5 using model (7.70)–(7.75), with (7.76) replacing (7.73) as in that example, with those from the LP model using only the third, horizontal segment of the CF as a capacity constraint. Figure 7.9 plots the total processing time required to process the demand for both products in the period in which it arises, against the maximum possible output of 20 units per period as a reference. It is apparent that for most of the planning horizon the system has considerable excess capacity, but will have to build anticipatory inventory to meet the demand peaks in periods 6 and 10.

The dual prices for capacity computed by the two models (λ_t for the ACF model and $\hat{\sigma}_t$ for the LP model) are shown in Fig. 7.10. The qualitative difference between the dual prices from the two models is immediately apparent. The LP model, which does not consider congestion, only returns positive dual prices for capacity in periods 6, 7, and 9, and these values are an order of magnitude lower than those for the ACF model. The dual prices for the ACF model begin increasing well ahead of the demand peaks representing the congestion caused by increasing releases, and reach substantial values even when the output of the system is below the theoretical maximum of 20 time units implied by the horizontal segment of the clearing function. The dual prices from the ACF are significantly higher than those for the LP model because they consider the additional workload required to raise output in the presence of congestion, which increases rapidly at high levels of output where the slope of the CF is small.

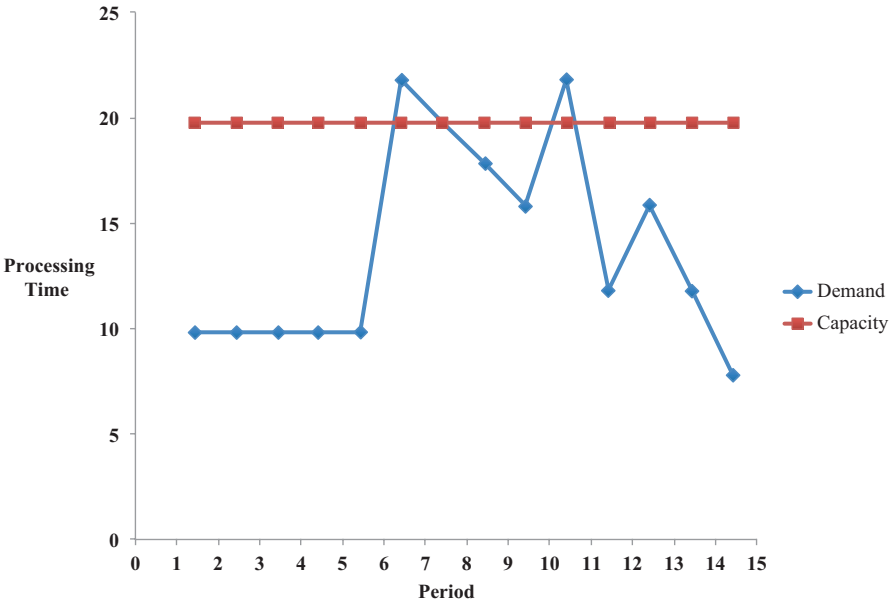

Fig. 7.9 Demand Levels for Examples 7.5 and 7.6

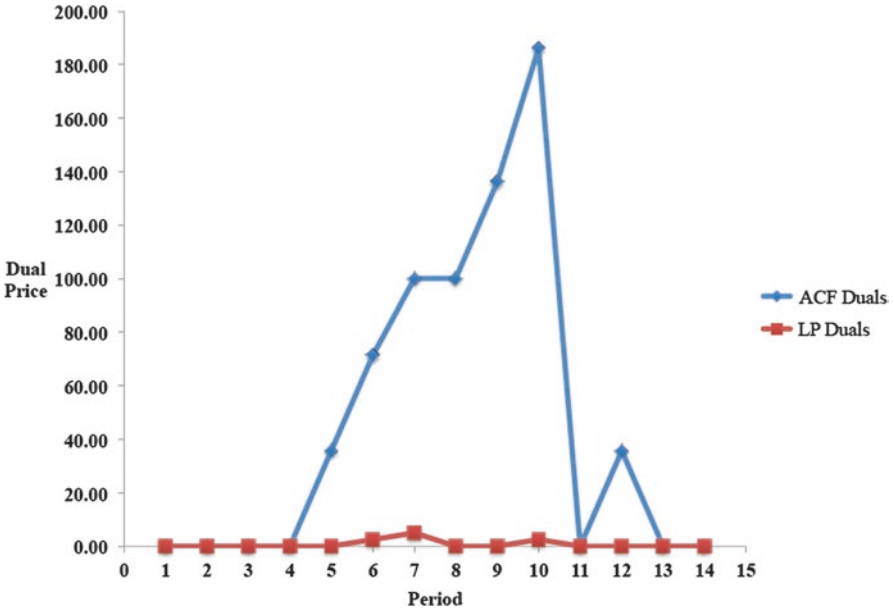

Fig. 7.10 Comparison of Dual Prices for LP and ACF Models in Examples 7.5 and 7.6

7.8 Conclusions

In this chapter, we have introduced the clearing function concept that provides a systematic approach to obtain tractable optimization models for release planning that recognize the nonlinear relation between workload, output, and cycle time discussed in the queueing models of Chap. 2. This chapter has focused on univariate clearing functions that represent the expected output of a production resource in a planning period as a concave non-decreasing function of a single state variable representing the amount of work available to the resource in the period. After reviewing several different types of clearing functions that adopt different state variables, we incorporate them into a convex optimization model for the single-product case. We then extend this model to illustrate the difficulties that arise in the presence of multiple products competing for capacity at a shared resource, and present the allocated clearing function formulation that provides an effective approximate solution to these difficulties. Finally, we show that the use of clearing functions leads to more informative dual prices for capacity than do the LP models of Chap. 5; in particular, the ACF model produces meaningful dual prices when resource utilization is below 1, which the LP models of Chap. 5 cannot.

While the clearing functions and the resulting optimization models described in this chapter have several desirable properties, especially those related to dual prices for resources and the more effective modeling of congestion, they also have some accompanying disadvantages. The need to include decision variables to explicitly

model WIP, and the piecewise linearization of the ACF model required to obtain an LP representation of this model, results in substantially larger formulations than those of Chap. 5. While the nonlinear form of the ACF model also yields a convex nonlinear program, due to the preservation of convexity by the perspective transformation, there is as yet little computational work exploring this area. Finally, the basic operation of the ACF model, which uses aggregate workload to estimate aggregate output and then allocates this aggregate output among competing products in a planning period, fails when this type of aggregation is no longer accurate, especially in the presence of significant setup times. In the next two chapters, we explore several more general clearing function models that seek to address these difficulties.

References

Asmundsson J, Rardin RL, Uzsoy R (2006) Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. IEEE Trans Semicond Manuf 19(1):95–111

Asmundsson J, Rardin RL, Turkseven CH, Uzsoy R (2009) Production planning with resources subject to congestion. Naval Res Logistics 56(2):142–157

Boyd S, Vandenberghe L (2009) Convex optimization. Cambridge University Press, Cambridge

Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice-Hall, Englewood Cliffs, NJ

Carey M (1987) Optimal time-varying flows on congested networks. Oper Res 35(1):58–69

Carey M (1992) Nonconvexity of the dynamic traffic assignment problem. Transport Res B 26B(2):127–133

Carey M, Bowers M (2012) A review of properties of flow–density functions. Transport Rev 32(1):49–73

Carey M, Subrahmanian E (2000a) An approach to modelling time-varying flows on congested networks. Transport Res B 34:157–183

Carey M, Subrahmanian E (2000b) An approach to modelling time-varying flows on congested networks. Transport Res Pt B Methodological 34(6):547

Curry GL, Feldman RM (2000) Manufacturing systems modelling and analysis. Springer, Berlin

Dafermos SC, Sparrow FT (1969) The traffic assignment problem for a general network. J Res Natl Bureau Standard B Math Sci 73B(2):91–118

Franklin RE (1961) The structure of a traffic shock wave. Civil Eng Publ Works Rev 56:1186–1188

Gopalswamy K, Uzsoy R (2018) Conic programming reformulations of production planning problems research report. Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University. Raleigh, NC

Gopalswamy K, Uzsoy R (2019) A data-driven iterative refinement approach for estimating clearing functions from simulation models of production systems. Int J Prod Res 57(19), 6013–6030.

Gopalswamy K, Fathi Y, Uzsoy R (2019) Valid inequalities for concave piecewise linear regression. Oper Res Lett 47:52–58

Graves SC (1986) A tactical planning model for a job shop. Oper Res 34(4):522–533

Graves SC (1988) Safety stocks in manufacturing systems. J Manuf Oper Manag 1:67–101

Graves SC, Kletter DB, Hetzel WB (1998) Dynamic model for requirements planning with application to supply chain optimization. Oper Res 46(3):35–49

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

Hannah LA, Dunson LA (2013) Multivariate convex regression with adaptive partitioning. J Mach Learn Res 14(1):3261–3294

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

Imamoto A, Tang B (2008) Optimal piecewise linear approximation of convex functions. World Congress on Engineering and Computer Science, San Francisco, CA

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

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

Karmarkar US (1989) Capacity loading and release planning with work-in-progress (WIP) and lead-times. J Manuf Oper Manag 2(1):105–123

Kefeli A, Uzsoy R (2016) Identifying potential bottlenecks in production systems using dual prices from a mathematical programming model. Int J Prod Res 54(7):2000–2018

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

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

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

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

Newell G (1961) A theory of traffic flow in tunnels. In: Herman R (ed) Theory of traffic flow. Elsevier, Amsterdam, pp 193–206

Nyhuis P, Wiendahl HP (2009) Fundamentals of production logistics: theory, tools and applications. Springer, Berlin

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. Netw Spat Econ 1(3-4):233–265

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

Srinivasan A, Carey M, Morton TE (1988) Resource pricing and aggregate scheduling in manufacturing systems. Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, PA

Teo C, Bhatnagar R, Graves SC (2011) Setting planned lead times for a make-to-order production system with master schedule smoothing. IIE Trans 43:399–414

Teo C, Bhatnagar R, Graves SC (2012) An application of master schedule smoothing and planned lead time control. Prod Oper Manag 21(2):211–223

Toriello A, Vielma JP (2012) Fitting piecewise linear continuous functions. Eur J Oper Res 219:86–95

Turkseven CH (2005) Computational evaluation of production planning formulations using clearing functions. School of Industrial Engineering, Purdue University, West Lafayette, IN

Van Aerde M, Rakha H (1995) Multivariate calibration of single regime speed–flow–density relationships. 6th Vehicle Navigation and Information Systems (VNIS) Conference, Seattle, WA

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