Chapter 8
Multivariate Clearing Functions

The clearing functions examined in Chap. 7 all assume that the expected output of a production resource in a planning period is a function of a single, aggregate state variable characterizing the amount of work available to the resource during the planning period; hence they were termed univariate clearing functions. As discussed in Chap. 7, several alternative definitions of this aggregate workload have been proposed, including the average WIP level during the planning period, the sum of entering WIP and new releases, or solely the beginning WIP. The use of such aggregate clearing functions in production environments with multiple products created anomalous behavior in the resulting optimization models as seen in Example 7.4. The allocated clearing function formulation develops an approximate formulation that provides effective solutions to this issue and has been validated in extensive computational experiments (Asmundsson et al. 2006, 2009; Kacar et al. 2012, 2013, 2016).

However, the allocated clearing function formulation is based on the assumption that the workload on the production resource resulting from all products in the system competing for its capacity can be aggregated into a single measure of workload without major loss of accuracy. An alternative statement of this assumption is that for a given total workload, however it is measured, the total amount of output, measured in the same units, that the resource can produce in a planning period is independent of the mix of products making up that total.

Univariate clearing functions also assume that the workload information for the current state, however defined, is sufficient to characterize the output of the resource in the current period. While this assumption may be valid for planning periods that are sufficiently long that the queues representing resource behavior can reach steady state and the periods of transient behavior at the beginning of the period due to new release decisions can be neglected, it is clearly questionable in many planning situations. Planning periods are often too short for steady state to be reached, and the release decisions introduced by the planning models at the start of each period are continually creating new workload situations by design. Queueing models suggest that the output of the system in any period can potentially depend on the entire

© Springer Science+Business Media, LLC, part of Springer Nature 2020

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_8

history of the arrival and service processes previous to the period, as well as their evolution during the period itself.

In this chapter, we shall examine more complex clearing functions that attempt to address these issues. The obvious first step is to disaggregate the single state variable for each period that forms the basis of the clearing functions in Chap. 7 in different ways. This approach begins by separating the two components of the period workload Λ_t into its two components, R_t and W_{t-1}, and treating each as a separate state variable. The presence of multiple products makes disaggregation of both WIP and releases by products a natural step. When cycle times exceed the length of the planning period, there may also be benefit to considering the workload in previous periods. For each set of state variables selected, a specific functional form for the clearing function must also be chosen. Many of these functional forms result in nonconvex optimization models, but there is considerable computational evidence that in many cases a standard convex solver yields high-quality solutions.

We shall begin our discussion by using transient queueing models to provide an initial intuition for why additional state variables are needed. We then discuss clearing functions that explicitly attempt to represent the transient behavior of the system without assuming steady state, and then proceed to consider additional state variables related to individual products and previous periods. The discussion of lot-sizing models based on multi-dimensional clearing functions that consider WIP levels, planned output levels, and planned lot sizes as state variables is treated separately in Chap. 9 since lot sizing raises some additional issues.

8.1 Limitations of Single-Dimensional Clearing Functions

The functional forms of the single-dimensional clearing functions described in Chap. 7 are almost all derived from steady-state queuing models. Hence they relate the average WIP or workload of the production system in steady state over a planning period to the expected output in this period. Similarly, a clearing function estimated from simulation data reflects the environmental conditions represented in the data set used to fit the clearing function. Any order release planning model using the clearing function thus implicitly assumes that these relationships continue to hold for each period of the planning horizon. However, since both demand and release quantities will vary over time, this assumption is often problematic. The order releases obtained from the clearing function model can exhibit characteristics that systematically deviate from steady state or from the characteristics of the simulation data used for setting the clearing function parameters, invalidating the shape of the clearing function assumed by the order release model.

This issue can be demonstrated by the following simple example. Consider a single production resource that can be modeled as an *M/M/1* queuing system in steady state. Recall from Chap. 2 that the clearing function for this system is given by (2.6). A clearing function based release planning model assumes that this function is valid for each period of the planning horizon. However, only the production

orders available to the resource at the start of period 1 are known with certainty since they can be observed directly. If the processing times are known with certainty, the initial WIP level W_0, measured in hours of work is thus known. If no releases of new work are expected during period 1, its workload will be $\Lambda_1 = W_0$, and the deterministic clearing function for period 1 will be

$$X_1 = \max\{\Lambda_1, C_1\} = \max\{W_0, C_1\} \tag{8.1}$$

unless machine breakdowns occur or work is delayed deliberately, which we shall assume is not the case. This is essentially the best-case clearing function of Hopp and Spearman (2008). Figure 8.1 compares the steady-state clearing function (7.24) derived by Missbauer for an *M/M/1* queue and (8.1). They clearly differ substantially, but a release planning model using a clearing function assumes that they are identical. In this case, the steady-state clearing function substantially and consistently underestimates the expected output of the resource in period 1 for a given workload, for the reasons discussed in (8.8) below.

This example describes an extreme case. We now generalize the underlying reasoning using the queueing-theoretical analysis presented below.

8.2 Transient Queueing Analysis of Clearing Functions

We arbitrarily select a particular planning period of an order release model, and consider a production resource modeled as an *M/M/1* queue that can be in transient regime. The number of the period is 1 without loss of generality, that is, the first planning period can have a negative period index. At the start of the period (time $t = 0$) the amount W_0 of WIP available to the resource, again measured in units of time, can be observed and hence is known with certainty. The resource is available for production for Δ time units during the planning period. We shall derive the

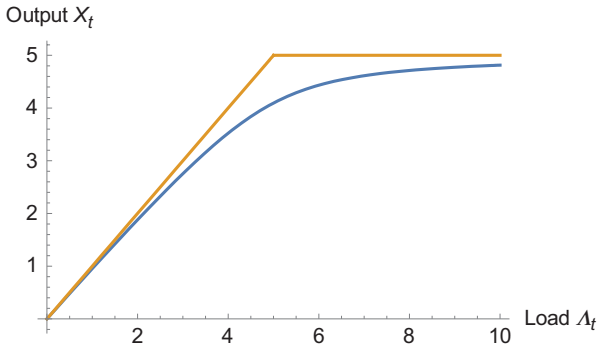

Fig. 8.1 Clearing function for idealized situation vs. for steady-state *M/M/1* system (Eq. (7.24) with $t_e = 0.2$, $\sigma = t_e$)

functional relationship between expected load $E[\Lambda_1]$ and expected output $E[X_1]$ of the resource for this period, where both quantities are measured in units of time. The analysis follows the approach of Missbauer (2011) where WIP is measured in number of orders; the following analysis for deterministic initial WIP is due to Missbauer (2014).

It is clear that if we have $W_0 \geq \Delta$, the output $X_1 = \Delta$. We now analyze the non-trivial case where $W_0 < \Delta$. In this case, the resource operates continuously from time $t = 0$ until time $t = W_0$. Within this time interval of length W_0 time units, work arrives according to a Poisson arrival process with arrival rate λ, but no arriving work is processed due to the FIFO assumption (highlighting, incidentally, the dependence of the clearing function on the specific dispatching policy used in the production unit). In contrast to the initial WIP W_0, which is known with certainty at time $t = 0$, we assume that no information is available about orders that arrive after the start of the period. Defining $p_n(t)$ as the probability of having n orders in the system at time t, the probability distribution of the number of orders in the system at time W_0, given by the number of orders arriving during the interval $[0, W_0]$, is

$$p_n\left(W_0\right) = \frac{\left(\lambda W_0\right)^n e^{-\lambda W_0}}{n!}, n = 0,1,\cdots \tag{8.2}$$

Given this probability distribution, the output of the system during the interval $[W_0, \Delta]$ can be derived by calculating the probability of idleness for all t in the interval $[W_0, \Delta]$ (Missbauer 2011). Denoting the output of the system in the time interval $[t_1,t_2]$ within period 1 by the random variable $X_1(t_1,t_2)$, the expected output in period 1 for mean arrival rate λ and initial WIP W_0 can be written as:

$$E\left[X_1\left(0,\Delta\right)|W_0\right] = \begin{cases} W_0 + E\left[X_1\left(W_0,\Delta\right)\right] & \text{for } W_0 < \Delta \\ \Delta & \text{for } W_0 \geq \Delta \end{cases} \tag{8.3}$$

We must now calculate $E[X_1(W_0, \Delta)]$, the expected output in the interval $[W_0, \Delta]$. After time $t = W_0$, the arrival process continues with rate λ until the end of the period. The state probabilities of having n orders in the system at time t, $W_0 < t \leq \Delta$, can be calculated from the state probabilities at time W_0 given by (8.2) and the conditional state probabilities $p_{rn}(t)$ of having n customers in the system at time t given r customers in the system at time 0. The latter is well-known in queueing theory (Cohen 1969: 82 ff. and 178) and is given by

$$\begin{aligned} p_{rn}\left(t\right) = {} & \left(1-u\right)\rho^n U\left(1-u\right) + u^{\frac{1}{2}(n-r)} e^{-(1+u)t/t_e} I_{n-r}\left(2\frac{t}{t_e}\sqrt{u}\right) - u^{\frac{1}{2}(n-r)} \\ & \int_t^{\infty} e^{-(1+u)\tau/t_e} \left\{ \begin{aligned} & I_{r+n}\left(2\frac{\tau}{t_e}\sqrt{u}\right) - 2u^{\frac{1}{2}} I_{r+n+1}\left(2\frac{\tau}{t_e}\sqrt{u}\right) + \\ & u I_{r+n+2}\left(2\frac{\tau}{t_e}\sqrt{u}\right) \end{aligned} \right\} \frac{d\tau}{t_e} \end{aligned} \tag{8.4}$$

for $t \geq 0$, where $I_j(x)$ denotes the modified Bessel function of the first kind, t_e the mean service time, $u = \lambda t_e$ the utilization and

$$U(t) = \begin{cases} 0, & t < 0 \\ 1/2, & t = 0 \\ 1 & t > 0 \end{cases}$$

The state probabilities at time $t > W_0$ are:

$$p_r(t) = \sum_{n=0}^{\infty} p_n(W_0)\, p_{nr}(t - W_0) \quad \text{for} \quad W_0 < t \leq \Delta \tag{8.5}$$

with $p_n(W_0)$ defined by (8.2). The expected output during the interval $[W_0, \Delta]$, measured in time units, is the expected total time during this interval the server is not idle:

$$E\left[X_1(W_0, \Delta)\right] = \Delta - W_0 - \int_{t=W_0}^{\Delta} p_0(t)\, dt \tag{8.6}$$

where $p_0(t)$ is obtained from (8.5) by setting $r = 0$. Substituting into (8.3) to calculate the output per period for deterministic initial WIP W_0, we obtain

$$E\left[X_1(0, \Delta) | W_0\right] = \begin{cases} \Delta - \int_{t=W_0}^{\Delta} \sum_{n=0}^{\infty} p_n(W_0)\, p_{n0}(t - W_0)\, dt & \text{for } W_0 < \Delta \\ \Delta & \text{for } W_0 \geq \Delta \end{cases} \tag{8.7}$$

Figure 8.2 illustrates the expected output (8.7) as a function of the expected workload in the period for different values of the initial WIP W_0. Missbauer (2011) presents the same analysis with initial WIP measured in number of orders. In that case, the differences in the expected output for different initial WIP levels are smaller because for a finite number of orders at the server at $t = 0$ there is always a positive probability of idleness within the period due to the exponentially distributed service times. Figure 8.2 clearly demonstrates that the entire shape of the clearing function changes based on the value of W_0, even when the latter is deterministic and not a random variable.

The assumption of deterministic initial WIP is reasonable for the first period in the planning horizon of an order release model. However, the initial WIP W_{t-1} available at the start of all subsequent planning periods t is a random variable. If we interpret the planned value of this random variable calculated in the release planning model as its expectation $E[W_{t-1}]$, the concavity of the clearing function and Jensen's inequality (Billingsley 1995: 80) yield

$$E\left[X_t(W_{t-1}, \Delta)\right] \leq E\left[X_t\left(E[W_{t-1}], \Delta\right)\right] \tag{8.8}$$

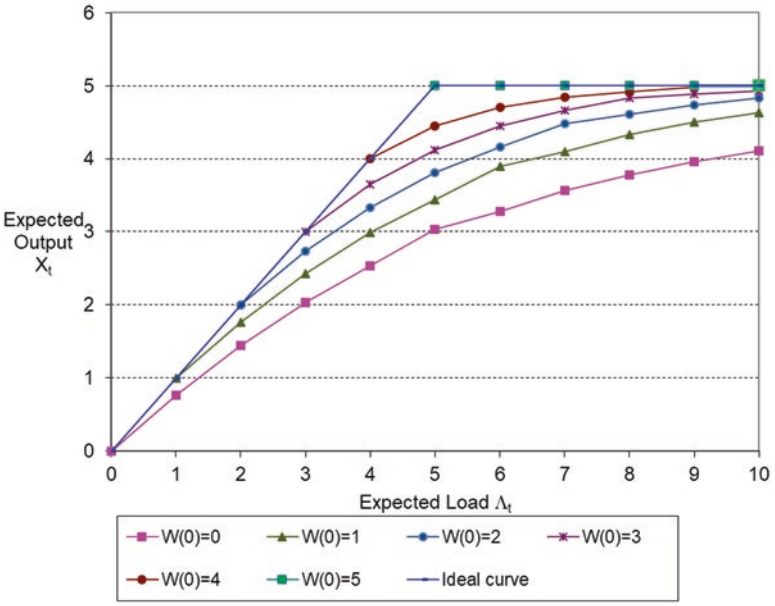

Fig. 8.2 Clearing functions for different deterministic initial WIP levels measured in hours of work. Period length $\Delta = 5$ time units, expected service time $t_e = 1$

implying that a clearing function treating the planned value of W_{t-1} as a deterministic parameter is likely to overestimate the expected output.

Continuing the analysis for period 1 with WIP measured in units of time, we define $f_{W_0}(w)$ as the probability density function of the initial WIP W_0. The expected output for given initial WIP W_0 is given by (8.7), and the expected output for stochastic initial WIP can then be obtained by conditioning as:

$$E[X_1] = \int_0^\infty E[X_1(0,\Delta)|w] f_{W_0}(w)\, dw \tag{8.9}$$

where $E[X_1(0,\Delta)|w]$ is given by (8.3).

Example 8.1 We consider the steady-state distribution of the initial WIP for the *M/M/1* system which, by the PASTA property that Poisson arrivals see time averages (Buzacott and Shanthikumar 1993: 54), is equal to the distribution of the (actual) waiting time of the arriving customers. This distribution is given by

$$f_{W_0}(w) = (1-u)\delta_0(w) + \lambda(1-u)e^{-1/t_e(1-u)w}, \quad \text{for} \quad w \geq 0 \tag{8.10}$$

where $\delta_0(w)$ denotes the Dirac Delta (unit impulse) function occurring at time $w = 0$ (Papadopoulos et al. 1993: 363).

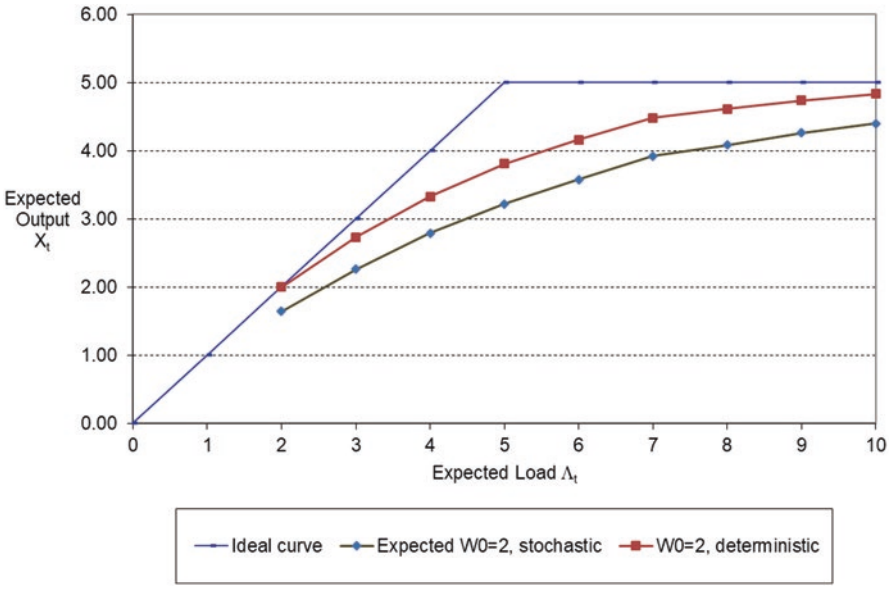

Fig. 8.3 Clearing functions for expected $W_0 = 2$, deterministic vs. steady-state distribution. $\Delta = 5$ time units, $t_e = 1$

The clearing functions (8.10) for different values of the *expected* initial WIP are plotted in Fig. 8.3 for the same data as in Fig. 8.2. Each point of $E[X_1]$ corresponds to a specified value of the arrival rate λ that, added to the expected initial WIP, leads to the expected workload given on the horizontal axis. For computational purposes, the numerical integral in (8.7) is discretized using 10 segments with a finite upper integration limit.

This analysis demonstrates that the expected output for a given *expected* load depends on the composition of the load (initial WIP vs. work released during the period), on the distribution of the initial WIP and also, implicitly, the probability distribution of the arriving work determined by the manner in which the new work is released over the duration of the period.

Armbruster et al. (2012) perform a similar analysis to that presented above for both a constant and time-varying arrival rate (influx, in their terminology) to the resource, analyzing the latter case using discrete-event simulation. They show that the functions depicted in Fig. 8.3 depend on the functional form of the influx over the period, concluding that "the clearing function cannot be just a parametric relationship between input and output" (p. 135).

Missbauer (2009) uses metamodels of the transient behavior of single-stage queueing systems developed from queueing models and simulation, specifically of the transient evolution of WIP over time, to estimate the output of a production resource per period. He shows that this leads to an integer, nonlinear formulation and that modeling errors occur that can lead to counterintuitive behavior. Hence at present the applicability of this approach is unclear.

The results of the analysis so far suggest that we may face a fundamental tradeoff in addressing the problem of formulating clearing functions. If one regards the clearing function as a metamodel of the behavior of the production resource of interest, the primary concern is to develop a model that best predicts the behavior of the resource for a given state at a given point of time. This suggests the use of sophisticated, high-dimensional statistical modeling methods such as Gaussian processes and time series analysis. Such techniques have been used by simulation researchers to develop the operating curves that describe expected cycle time as a function of resource utilization (Yang et al. 2006; Ankenman et al. 2010). Li et al. (2016) use similar techniques to develop a metamodel predicting the output of a production system over time based on a number of state variables, which they then use in place of a discrete-event simulation model in a simulation optimization approach.

While properly formulated and calibrated models of this kind are capable of predicting the output of a production resource or production unit in a planning period quite accurately, they are generally unsuitable for use in a mathematical programming model due to their complex functional forms. As we shall see later in this chapter, even relatively simple multivariate clearing functions lead to non-convex order release models. Hence there appears to be a basic tradeoff between computational tractability of the resulting order release model and the accuracy of the output estimates produced by a clearing function. This issue will surface frequently in the discussion of different functional forms for multivariate clearing functions in this chapter.

Selçuk et al. (2008) formulate a "short-term nonlinear" (STN) clearing function assuming that the WIP at the server is measured in number of orders. Each order that contributes to the load in a certain period is available as soon as it is needed for processing. For exponentially distributed service times, the departure process from the server is a Poisson process with mean rate equal to the service rate μ until the last order available in this period is completed, after which the server is idle. Under these assumptions, the expected output as a function of the number of available orders (i.e., the workload) can be calculated. Note that idle time at the server due to stochastic interarrival times cannot occur in this model. This simplification allows the univariate clearing function to model the transient state. The saturating shape of this CF is due to the uncertain work content of the orders, which is assumed to be unobservable even for the initial WIP at the time of planning. Asmundsson et al. (2009) use a similar but somewhat more general formulation to prove the concavity of the clearing function in a transient regime.

An approximate model of transient queuing systems that can be integrated into order release models is the stationary backlog carryover (SBC) approach introduced by Stolletz (2008) for *M(t)/M(t)/c(t)* systems and extended to *G(t)/G/1/K* systems by Stolletz and Lagershausen (2013). We shall describe the technique for an *M(t)/M/c* system, characterized by a time-varying Poisson arrival process, exponential service times, and c servers. In the SBC approach, time is divided into short intervals, usually equal in length to the mean service time t_e, with arrival rate λ_t during each interval t. We shall refer to these short intervals as micro-periods to distinguish them from the longer planning periods discussed throughout the volume. The average

utilization in period 1 is assumed to be equal to the steady-state utilization of an *M/M/c/c* queue with arrival rate λ_t, which is given by

$$E[u_1] = \lambda_1 g(\lambda_1) t_e \tag{8.11}$$

where $g(\lambda)$ denotes the steady-state fraction of served customers in an Erlang loss (*M/M/c/c*) system with c servers and a mean service time t_e as a function of the arrival rate λ. Recall that a finite capacity queue or loss system with capacity c can accommodate at most c customers; an arriving customer encountering c customers already in the system will depart without being served. Hence, in this model a fraction $P_1 = 1 - g(\lambda_1)$ of the arriving orders will be blocked from entering the system, giving the expected number of blocked orders in period 1 as

$$b_1 = \lambda_1 P_1 \tag{8.12}$$

For all subsequent micro-periods $t = 2, 3,\ldots$, an artificial arrival rate $\tilde{\lambda}_t$ that accounts for both the (artificial) backlog and new external arrivals is calculated as

$$\tilde{\lambda}_t = b_{t-1} + \lambda_t, \quad t = 2,3,\ldots \tag{8.13}$$

The expected utilization is then calculated from this artificial arrival rate as

$$E[u_t] = \tilde{\lambda}_t \; g(\tilde{\lambda}_t) t_e, \quad t = 2,3,\cdots \tag{8.14}$$

Note that if the output estimate is correct the expected artificial backlog b_{t-1} represents the expected WIP, measured in number of orders, in the real system at the end of micro-period $t - 1$ and hence the start of micro-period t. Hence (8.14) calculates the expected utilization as a concave saturating function of the workload λ_t, making SBC a special case of a clearing function model (Missbauer 2007), but with an equality constraint on the output. Missbauer and Stolletz (2016) formulate and test an order release model based on SBC, finding it to be mathematically consistent and solvable by standard NLP solvers. Closely related approximate queueing models for transient systems are suggested by Askin and Hanumantha (2018).

8.3 Transient Clearing Functions with Multiple Variables

The problems with the usual one-dimensional clearing functions are obvious: they express the relationship between load and output under long-run average (steady-state) conditions although the actual relation is conditional on the history prior to the planning period under consideration and can be very different from any steady-state condition, especially for the first period of the planning model where initial WIP is largely deterministic. This suggests that extending the one-dimensional CF

with additional explanatory variables that reflect the history of the process should improve its ability to estimate output. The queueing-theoretical results derived above indicate that disaggregating the period workload into initial WIP and releases during the period is the most obvious extension. This leads to a two-dimensional CF of the form

$$X_t = f\left(W_{t-1}, R_t\right) \tag{8.15}$$

where R_t denotes the work input in period t.

Andersson et al. (1981) propose a linear clearing function of this form without explicit reference to the queueing argument above. In our context, a saturating clearing function of the form (8.15) must be used in order to reflect the congestion phenomena arising from the stochastic nature of arrivals and service times and the finite capacity of the resource. Deriving a piecewise linear approximation of such a function based on empirical or simulated data is difficult without postulating an underlying nonlinear functional form. In contrast to steady-state queueing models, the expressions describing the behavior of transient queueing systems can only be evaluated numerically, rendering the derivation of a tractable expression for a saturating, two-dimensional clearing function difficult. Häussler and Missbauer (2014) propose what appears to be a reasonable functional form with the following properties:

- A fraction β of the initial WIP, measured in time units (hours of work), is converted into output during the period, up to the maximum available capacity C_t. Simulation models generally assume $\beta = 1$. In general, β is a parameter whose value must be estimated from the data.
- For positive releases $R_t > 0$ the output $X_t \leq Min\ \{W_{t-1} + R_t, C_t\}$.
- For given initial WIP W_{t-1} the clearing function is concave and monotonically non-decreasing (saturating) in R_t, with

$$\lim_{R_t \to \infty} f\left(W_{t-1}, R_t\right) = C_t \tag{8.16}$$

- For $R_t > 0$ we assume that for fixed W_{t-1} the increase of the clearing function with R_t follows the same functional form as the one-dimensional clearing function derived from a steady-state *M/G/1* model in (7.24). The two-dimensional clearing function is then

$$X_t = \begin{cases} \beta W_{t-1} + \dfrac{\left(C_t - \beta W_{t-1}\right)}{C_t} \cdot \dfrac{1}{2}\left[C_t + k + R_t - \sqrt{C_t^2 + 2C_t k + k^2 - 2C_t R_t + 2kR_t + R_t^2}\right] \\ \qquad\qquad \text{if } W_{t-1} < C_t / \beta \\ C_t \qquad \text{if } W_{t-1} \geq C_t / \beta \end{cases} \tag{8.17}$$

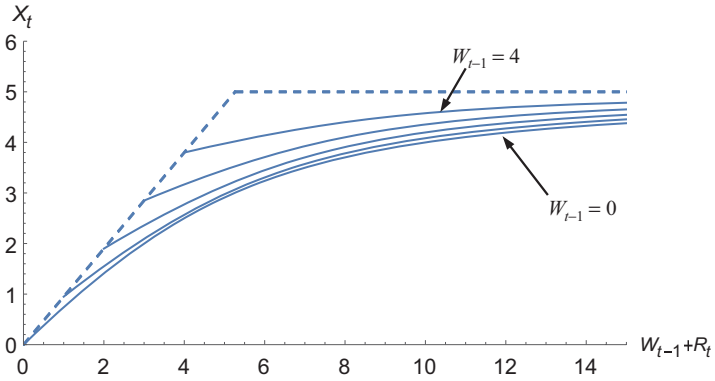

Fig. 8.4 Two-dimensional CF (Equation (8.17)) for initial WIP levels from 0 (lower function) to 4. Parameters: $C_t = 5$, $k = 1.5$, $\beta = 0.95$. The dashed line is the ideal curve $Min\{W_{t-1} + R_t, C_t\}$

Figure 8.4 depicts (8.17) and shows that, when parameterized appropriately, this CF exhibits a shape very similar to that of the transient CF in Fig. 8.2.

This logic suggests that the fit of the CF can be improved by switching from a 1-dimensional to a 2-dimensional CF. This hypothesis is tested in Häussler and Missbauer (2014) for both simulated and empirical data. The quality of the fit is measured by the adjusted coefficient of determination (adj. R^2). Overall, the hypothesis is confirmed for bottleneck resources although the improvements in fit are often smaller than one might expect. To the best of our knowledge no significance test for changes in R^2 exists, but the sample size is large (1690 periods for simulation, 350 periods for the empirical data).

Table 8.1 shows the R^2 values for three machines operating at the manufacturer of optical storage media described in Chap. 1; the simulation represents a scaled-down version of this manufacturing system. Note the substantial difference between the results for simulated and empirical data caused by the noise in the empirical data, as also observed by Fine and Graves (1989). As expected, the fit for simulation data depends on the period length. In Table 8.1, a period length of five times the average processing time t_e is used. For the period length of the empirical data, which is about 15 times the average processing time, the adj. R^2 for simulation is very close to 1.

A saturating, 2-dimensional clearing function based on W_{t-1} and R_t such as (8.17) leads to a convex, nonlinear order release model. Although there is little experience with this structure, it appears to be computationally tractable. Approximating (8.17) by a set of linear functions, in the manner used for 1-dimensional clearing functions, leads to a high number of constraints in the resulting LP model. Successive linear approximation in the optimal region (Hadley 1964) is an alternative as well as using NLP solvers. Determining the best way to solve the resulting models remains a topic for future research.

Table 8.1 Adjusted R^2 for representative bottleneck machines in the manufacturing (Man), printing (Pri), and packing (Pack) department of an optical storage media manufacturer

Machine	Utilization	R^2 1-dim. CF	R^2 2-dim. CF
Simulation data			
ManBNS	Gateway workcenter		
PriBNS	95.34%	0.743	0.939
PackBNS	71.95%	0.937	0.977
Empirical data			
ManBN	88.71%	0.664	0.687
PriBN	82.77%	0.578	0.600
PackBN	80.03%	0.656	0.702

1-dim. CF Equation (7.24), 2-dim. CF Equation (8.17) (Häussler and Missbauer 2014)

Decomposing W_{t-1} into its components W_{t-2} and R_{t-1}, which leads to a three- or more dimensional clearing function with explanatory variables that reflect the evolution of work input and output over time, has not been considered in the research so far. Kacar and Uzsoy (2014) explore this issue using a product-based clearing function that makes no distinction between the different operations l of each product at each workcenter, but fits a clearing function for each product at each workcenter. Thus the release, WIP, and output variables are defined as

$$R_{ikt} = \sum_{\{l:k=k(il)\}} R_{itl}, W_{ikt} = \sum_{\{l:k=k(il)\}} W_{itl}, X_{ikt} = \sum_{\{l:k=k(il)\}} X_{itl} \tag{8.18}$$

and clearing functions $f_{ik}(.)$ are formulated for each product $i \in I$ and workcenter k. Plotting the total output for all products against the total initial WIP $\sum_{i\in I} W_{ikt}$ and total releases $\sum_{i\in I} R_{ikt}$ to a workcenter k subject to machine failures, as illustrated in Fig. 8.5, suggests that there is benefit in disaggregating the workload into its components, releases R_t and entering WIP W_{t-1}, as suggested in (8.15). Hence they propose three different product-based clearing functions. Model 1 uses only state information for the current period, given by

$$X_{ikt} = \mu_{ik} + \sum_{i\in I} \beta_{ik} R_{ikt} + \sum_{i\in I} \theta_{ik} W_{ik,t-1} \tag{8.19}$$

where μ_{ik} denotes the intercept and β_{ik} and θ_{ik} the regression coefficients to be estimated. Model 2 extends Model 1 by considering the releases of product g in the immediately preceding period, yielding

$$X_{ikt} = \mu_{ik} + \sum_{i\in I} \beta_{ik} R_{ikt} + \sum_{i\in I} \theta_{ik} W_{ik,t-1} + \psi_{ik} R_{ik,t-1} \tag{8.20}$$

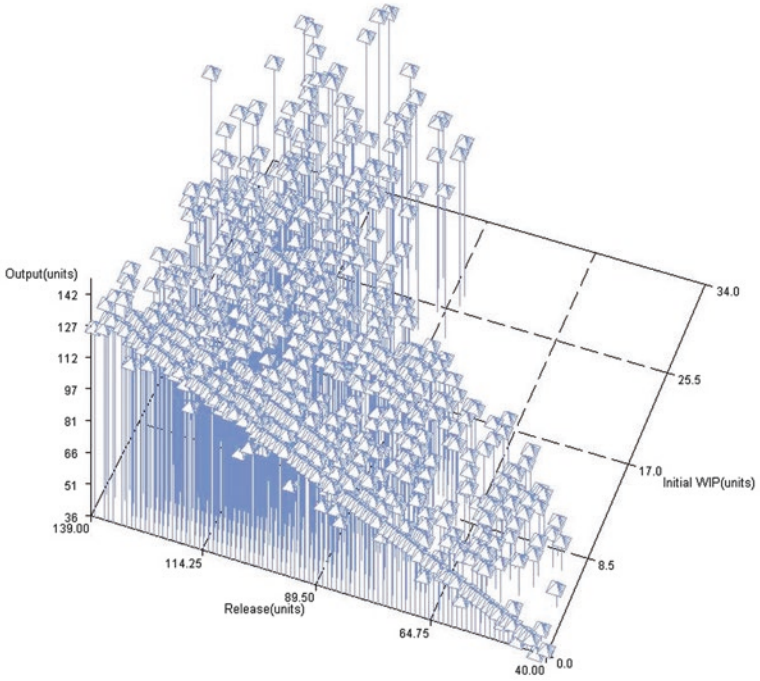

Fig. 8.5 Total Output at a Machine Subject to Failures as a Function of Releases and Initial WIP (Kacar and Uzsoy 2014)

The final model, Model 3, augments Model 2 by adding the releases for all products in the immediately preceding period, giving

$$X_{ikt} = \mu_{ik} + \sum_{i \in I} \beta_{ik} R_{ikt} + \sum_{i \in I} \theta_{ik} W_{ik,t-1} + \sum_{i \in I} \psi_{ik} R_{ik,t-1} \tag{8.21}$$

The planning model using the product-based clearing functions differs somewhat from the ACF model used with the workload-based clearing function discussed in Sect. 7.2.3. The objective function remains the same, assuming all operations of a given product incur the same WIP holding cost. The material balance equations for finished goods inventory of each product and WIP of each operation of each product are also the same as in the ACF model. However, the constraints governing the output of each product in each period are given by

$$X_{ikt} \leq f_{kt}(.), i \in I, k \in K, t = 1, \ldots, T \tag{8.22}$$

$$\sum_{i \in I} X_{ikt} \leq C_k, k \in K, t = 1, \ldots, T \tag{8.23}$$

where $f_{kt}(.)$ is defined by one of (8.19), (8.20), or (8.21). Constraints (8.23) were included because, occasionally, the fitting procedure will return a fit whose intercept exceeds the theoretical capacity of the workcenter.

The comparison of the different product-based clearing functions sheds some light on the issue of whether or not to include state variables related to previous history in the clearing function. Under low utilization the clearing functions (8.19) that consider only variables for the current period are among the best performers, although the difference in expected profit between the models is sometimes small (though statistically significant). At high utilization the model (8.20) that includes the releases of the individual product from the previous period is consistently among the best performers. These results are in general intuitive: at lower utilization levels the production resources will be able to convert the majority of the available workload in a period into output, leaving little WIP at the workcenter at the end of the period. When utilization increases, cycle times will also increase, causing the releases in the previous period to affect output in the current period.

An interesting finding of this work is the analysis of the residuals from the regression models fitted. Figure 8.6 shows the residuals (difference between predicted and realized output) of one of the product-based clearing functions as a function of the observed output of one of the products. Ordinary least-squares regression assumes that the residuals should be independent and normally distributed with homogenous variance and mean zero. It is apparent from Fig. 8.6 that the model illustrated did

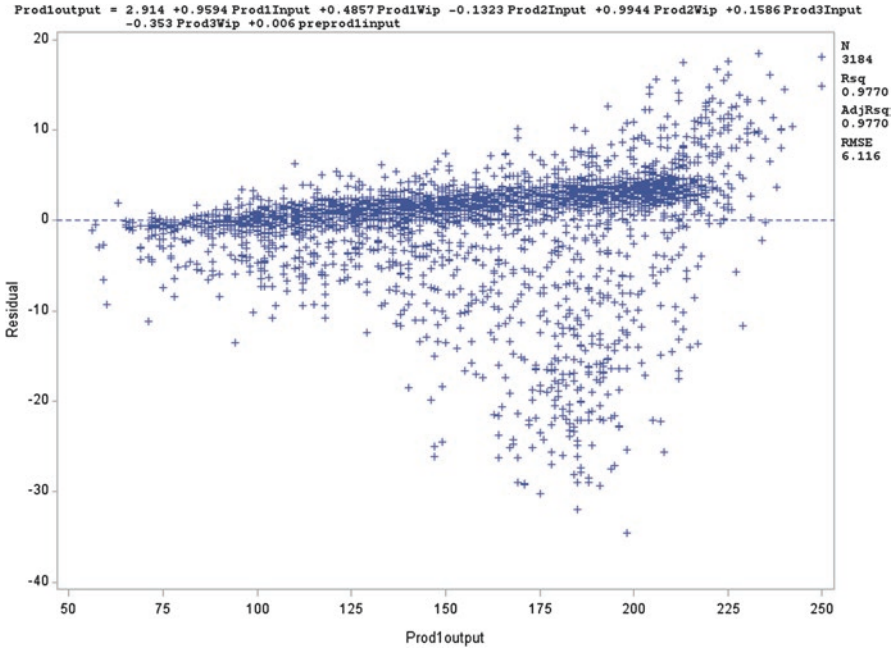

Fig. 8.6 Residuals for Product-based Clearing Function (8.20) of Unreliable Machine

not satisfy these conditions. While the mean residual is close to zero at low output levels, as output levels increase an upward trend appears. In addition, the variance of the residuals for a given output level, shown in the figure by the vertical dispersion of the points around the horizontal axis, is also increasing, and far from symmetric, suggesting frequent underestimation in the output range 120–220 units. At very high output levels, the problem seems to be one of systematic overestimation. Clearly, the interactions between the state variables are complex, and there is much room for improvement.

In hindsight, the failure to distinguish between the workloads of different operations, i.e., workload of the same product at different stages of processing, confounds the results of these experiments considerably. Comparison of the product-based clearing function and the workload-based clearing functions used in the ACF model shows, unsurprisingly, that in five of the eight experimental conditions the workload-based clearing function outperforms the various product-based clearing functions. The product-based clearing functions perform better for both low utilization short failure cases and low utilization long failures with high demand CV. The reason for this lies in the more granular representation of the production resources in the workload-based clearing function. Recall that in the product-based clearing functions there is no information capturing the flow of material through the different operations of each product routing; the product-based clearing function considers only the total number of lots of each product processed in that period. This creates the opportunity for incorrect behavior such as that illustrated in Chap. 7 for single-variable clearing functions. The product-based clearing function for a given product must produce the different operations in the right combination, but there is nothing in the model to ensure this apart from the finished goods inventory balance equation, which meets demand for each product using output from the last operation on its routing. In contrast, the workload-based clearing function creates a single clearing function for the workcenter whose capacity is shared among the operations, and uses the allocated clearing function formulation to allocate the estimated total output of the workcenter among all operations of all products processed there. The observation that the workload-based clearing function outperforms the best product-based clearing functions in five of the eight experimental conditions, particularly those at high utilization, suggests that the product-based clearing functions as implemented in this study are deficient in multiple aspects. The results of Albey et al. (2014, 2017) discussed below, which examine different aggregations of state variables in single- and multistage production systems, also contribute to this discussion.

Häussler and Missbauer (2014) examine the fit of various 3-dimensional clearing functions to the empirical and simulated data for the manufacturer of optical media described in Chap. 1 and for simulation data specifically designed for this experiment. Since no functional form for saturating 3-dimensional clearing functions is known, they use a linear and a specific cubic function. Although minor improvements in fit were observed in some cases, the results are largely inconclusive. This suggests rapidly diminishing returns to increasing the dimensionality of the clearing functions, but this must be examined in further studies.

The findings presented so far demonstrate that the expected output in a given planning period depends, in principle, on the entire history of the process up to the current period. Neglecting this dependence leads to an inaccurate estimate of the expected output in the planning period, which can be termed an *estimation error*. The inclusion of this inaccurate clearing function in the order release model leads to suboptimal releases over time, which we shall term *optimization error*. In particular, since the clearing function represents the expected output of the system for a given state and time as opposed to its maximum possible output, the effects of temporary periods of high workload (workload peaks) are unlikely to be predicted accurately. A number of experiments have shown that CF-based order release models can lead to fluctuations in releases over time that exceed those in external demand (Missbauer 1998, 2009; Bischoff 2017), which might well be due to this estimation error. Orcun and Uzsoy (2011) observe oscillations of this type in a system where the planning model assumes a fixed lead time but realization follows a clearing function, creating a mismatch between the planning model and the system it is representing.

However, the relationship between the fit of the CF and the quality of the release schedules (at the discrete-event level) is complex (Kacar and Uzsoy (2015)). Preliminary numerical experiments with 2-dimensional CFs show that they can lead to high variations of the releases over time. Figure 8.7 depicts the optimization results for the single-stage, single-product CF model described in Sect. 7.2 (repeated for convenience) that seeks to minimize

$$\min \sum_t wW_t + \sum_t hI_t \tag{8.24}$$

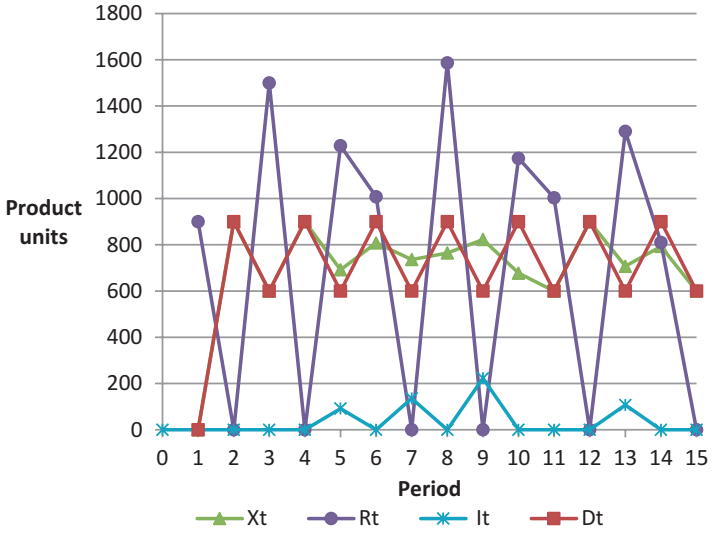

Fig. 8.7 Optimization result for oscillating demand

subject to the standard WIP and finished goods inventory balance equations, the CF (8.17) with $\beta = 1$ and nonnegativity constraints for all variables. The WIP holding cost coefficient $w = 0.5$, and the FGI holding cost $h = 1$ per unit-period. The parameters of the CF are $k = 200$, $C = 950$.

This behavior appears to arise because output in a given period can be increased by either providing initial WIP or by releasing work in the period. Providing initial WIP generates capacity more efficiently since all of it is cleared up to the available capacity. Releasing new work generates less capacity due to the nonlinearity of the CF in R_t. For instance, in Fig. 8.7 900 units are released in period 1, held back ($W_1 = 900$) and processed in period 2 ($X_2 = 900$) since this is cheaper than releasing more work in period 2 in order to generate a capacity of 900. This point at which it becomes more economical to release WIP rather than hold it back will change with the utilization due to the specific nonlinear shape of the CF (8.17) that is depicted as a contour plot in Fig. 8.8. This is also related to the findings of Carey (1987) that holding back behavior will arise when releasing WIP in the current period will cause congestion in later periods. This is counterintuitive and indicates that integrating the history of the process into order release models requires modification of the model structure as well. How to do this is largely a topic for future research.

The fact that the expected output in a period depends on the entire history of the process up to that period leads to another important issue: Except for initial WIP of the order release model, the values of the independent variables of the clearing function are point forecasts of a future state of the system, and hence subject to random forecast error, which influences the expected output as seen in Fig. 8.3; different realized values of the initial WIP result in a different curvature for the clearing function. Describing this forecast error for some future period as a function of the deci-

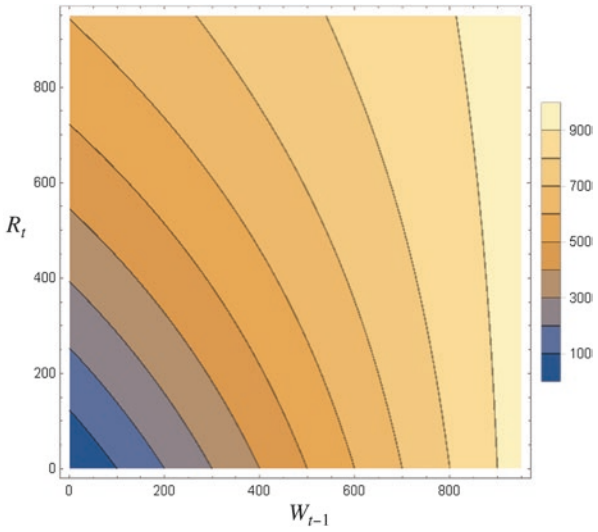

Fig. 8.8 Contour plot of the CF (8.17) for $\beta = 1$, $C_t = 950$, $k = 200$

sion variables in a release planning model is difficult for two reasons. Firstly, it is not based on hard data that can be measured, but instead reflects the decision maker's state of knowledge at a certain time, e.g., the accuracy with which the WIP level on Thursday morning can be estimated on Monday morning, given specified work releases during the intervening period. Stochastic models of the evolution of the forecast error over time are required. The Martingale Model of Forecast Evolution (Heath and Jackson 1994) is one such approach that has been successfully applied to production planning under uncertain demand (Albey et al. 2015). Secondly, errors in the WIP estimation will increase as the future periods become more remote. Integrating these factors into the order release model results in a complex stochastic programming problem since the evolution of information over time must be considered in a rolling horizon planning framework (Missbauer 2014). While some initial efforts have been made to formulate stochastic optimization models of such problems (Aouam and Uzsoy 2012, 2015; Albey et al. 2015; Lin and Uzsoy 2016), the development of scalable, practically applicable models remains a topic for future research.

8.4 Multivariate Multiproduct Clearing Functions

The second principal motivation for the development of multivariate clearing functions is the need to consider the interactions of multiple products competing for capacity at the production resources of interest. This issue has already raised its ugly head in Chap. 7, where we saw that when a univariate clearing function based on a state variable aggregated over different products is used, counterintuitive behavior can result even in the absence of setup times between products. The allocated clearing function formulation addresses this issue to a degree of approximation in the absence of significant setup times. We shall show in this section that when contention between multiple products can lead to significant loss of output, as is the case in the presence of setups, the univariate clearing function fails to predict output at the level of individual products.

We shall first use a simple aggregate queueing model to explore the impact of multiple products on the output of a production resource. We then examine a number of multivariate clearing functions that explicitly address the presence of multiple products, and then consider production units with internal routing flexibility. Under these conditions it is no longer possible to describe the behavior of the production resources using a single clearing function; instead, a system of nonlinear clearing functions that describe the output of each item for fixed WIP and output levels of all other products in the system is required.

8.4.1 Motivation

The simple, steady-state queueing analysis used in Chap. 2 can be extended to examine the impact of product mix on system output. In that chapter, we had shown that the average utilization u of a *G/G/1* queue in steady state as a function of the average WIP level W can be approximated as

$$u = \frac{-(W+1)+\sqrt{(W+1)^2+4(\Psi-1)W}}{2(\Psi-1)}, \quad \text{for} \quad \Psi \neq 1 \tag{8.25}$$

where $\Psi = \left(c_a^2 + c_e^2\right)/2$, c_a^2 denotes the squared coefficient of variation of the interarrival times and c_e^2 that of the effective service time. Recall that the effective service time is a random variable representing the amount of time a job will spend in service, taking into account both the natural processing time and disruptions such as setups, quality issues, and machine failures (Hopp and Spearman (2008), Chap. 8).

If significant setup times must be incurred when switching between different products, the impact of product mix on the distribution of the effective processing time can be characterized as in Hopp and Spearman (2008). Suppose the natural processing time, the time required to process a job without any detractors such as setups and machine failures, has mean t_0 and variance σ_0^2. Assuming a setup is equally likely to occur after each part being processed, with the average number of parts processed between setups being N_s, and denoting the mean and variance of the setup time by t_s and σ_s^2, respectively, the mean and variance of the effective processing time are given by Hopp and Spearman (2008), Chap. 8:

$$t_e = t_0 + \frac{t_s}{N_s} \tag{8.26}$$

$$\sigma_e^2 = \sigma_0^2 + \frac{\sigma_s^2}{N_s} + \left(\frac{N_s - 1}{N_s^2}\right) t_s^2 \tag{8.27}$$

The mix of products processed by the system can potentially affect all terms in the expressions above. The more frequently setups need to be performed, the smaller N_s will be; in addition, both the mean and variance of the setup time distribution may increase as a more diverse portfolio of products requiring different equipment configurations are processed. In practice, lot sizing policies will affect N_s, and continuous improvement programs such as single minute exchange of die (SMED) (Shingo 1986) seek to reduce both t_s and σ_s^2. However, the impact of product mix on utilization, and hence output, through its impact on c_e^2 is evident.

A simple simulation experiment reported by Albey et al. (2014) makes this point quite graphically (no pun intended!). They consider a single-stage production system capable of producing two different parts, whose processing times are lognormally distributed with a mean of t_0 = 100 s and a coefficient of variation of 0.13.

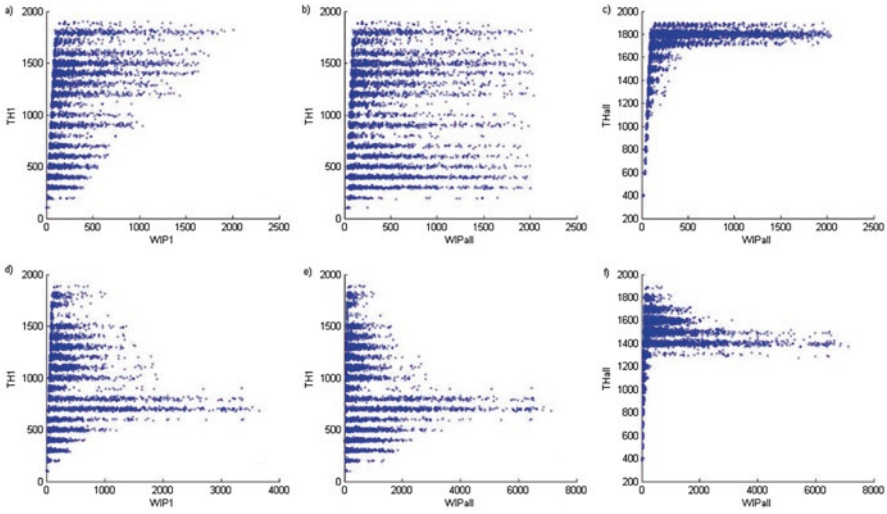

Fig. 8.9 Impact of Product Mix on System Output

Parts are released into the system one by one following a cyclic pattern based on the heuristic of Askin and Standridge (1993). They consider two different situations: one in which there are no setup times between part types, and one where the setup time follows a triangular distribution with mean $t_s = 0.1t_0$. The demand in each period follows a Poisson distribution, leading to a mean total workload of 1600 s in each period. The total workload for the period is then disaggregated into individual products over 10 different product mixes, where the ratio of the second product to the first ranges from 0 to 5 (i.e., 0, 0.2, 0.25, 0.33, 0.5, 1, 2, 3, 4, and 5). Each product mix is simulated for 1000 different workload realizations, resulting in a total of 10,000 observations of workload and output. The resulting plot of the output of the system in a planning period of 1800 s is shown in Fig. 8.9.

The upper row of graphs represent the performance of the system without setup times. The first two graphs plot the output of Product 1 in the planning period as a function of the WIP of Product 1 and the total WIP in the system, in units of time; the rightmost graph shows total output of both products as a function of the total WIP of both products. The banded appearance of the two leftmost charts is due to the discrete product mix combinations used in the experiment. A specified output of Product 1 can be obtained for various WIP levels of that product (leftmost graph), or of all products (middle graph), depending on the amount of Product 2 in the system. Hence the output of Product 1 is not well described by either its own WIP or the total WIP of both products. The rightmost chart, however, shows that the total system output of both products is well represented by a function of the total WIP.

The lower panel of graphs tells a similar story—representing the output of Product 1 in terms of a WIP measure is inaccurate. However, in the presence of setup times, the output of Product 1 can decrease as its WIP increases, if the amount of Product 2 in the system is also increasing. For a given level of either WIP mea-

sure (Product 1's WIP or the total WIP), different output levels of Product 1 can be achieved depending on the amount of Product 2 in the system. The rightmost graph in the lower row differs qualitatively from that above it, showing that in the presence of setup times the aggregate output of the system does not present a monotonically increasing, concave shape.

Motivated by these observations, Albey et al. (2014) examine a number of different multi-dimensional clearing functions (MDCFs) for a single production resource. Their point of departure is the univariate clearing function of Karmarkar (1989), given by

$$f\left(\Lambda_t\right)=\frac{K_1\Lambda_t}{K_2+\Lambda_t},\quad \Lambda_t\geq 0 \tag{8.28}$$

where Λ_t denotes the workload available to the resource throughout period t as discussed in Chap. 7. They note that in a multiproduct environment, the output of a given product in a planning period must depend on both the amount of that particular product available to the resource during the period, and the amount of capacity allocated to other products. The allocated clearing function formulation of Chap. 7 addresses this issue by estimating the aggregate output of the resource in units of time as a function of the total workload of all products available to it, and then disaggregating this into estimates of output for individual products. Albey et al. (2014) take a different approach by formulating a MDCF for each product i, representing the capabilities of the resource by a system of nonlinear, linked clearing functions that use state variables related to all products in the system in the planning period. They consider two classes of these MDCFs: WIP-based MDCFs (W-MDCFs), where the impact of other products $j\neq i$ in the system is represented by the average WIP level of each product during the planning period; and output-based MDCFs (O-MDCFs), where the impact of the other products is estimated using their planned output. They experiment with several functional forms of each type, represented by the O-MDCF

$$X_i=\frac{\left(C-\sum_{j\neq i}a_jX_j\right)\bar{W}_i}{M_i-b_i\sum_{j\neq i}a_jX_j+\bar{W}_i} \tag{8.29}$$

where X_i denotes the expected output of product i in the planning period, $\bar{W}_i$ the planned time-average WIP level of product i over the period, and M_i, a_i and b_i are user-defined parameters to be estimated from data. C denotes the expected capacity of the resource in the period. The general form of the W-MDCFs is

$$X_i=\frac{a_i\bar{W}_i+b_i\sum_{j\neq i}\bar{W}_j}{M_i+\sum_j b_j\bar{W}_j} \tag{8.30}$$

Several different versions of each MDCF family, the details of which are given in Albey et al. (2014), were tested in computational experiments. The authors show that the MDCFs are non-convex functions, so that the resulting release planning models can be reduced to quadratically constrained nonlinear programs (Linderoth 2005; Bao et al. 2011), which are known to be strongly NP-hard but can be solved by enumerative methods using solvers such as BARON (Tawarmalani and Sahinidis 2005). Some specific MDCFs belong to the class of bilinearly constrained bilinear problems (Al-Khayyal 1992), whose non-convex nature appears to be less severe than that of the general quadratically constrained nonlinear problem. The nine different MDCFs are fitted using least-squares regression using an extensive set of training data generated from a simulation model of a resource processing four different products in different proportions. They consider three different experimental situations. In the first, there is no loss of capacity in switching from one product to another, and products are processed in FIFO order. In the second case switching from one product to another involves a tool change time, with FIFO dispatching. The final case assumes no tool change time and dispatching in order of Shortest Processing Time (SPT), to examine the impact of shop-floor dispatching policy on the performance of the various MDCFs. In all experiments, the products to be released in a period are sequenced in a cyclic pattern and released all together at the start of the planning period, which will result in a very large number of tool changes in the second experimental configuration. The performance of the MDCFs is measured by implementing them in a release planning model, consisting of balance equations for the WIP and finished goods inventory of each product and the MDCFs for each product in each period, and simulating the performance of the production system under the releases determined by these models. Since obtaining globally optimal solutions to the resulting non-convex optimization models requires very high CPU times, the authors use a convex nonlinear solver to obtain locally optimal solutions.

Under FIFO dispatching without tool changes, all but the most simplistic MDCFs perform comparably with the ACF model and a simpler LP model that ignores congestion. The striking feature of this experiment is the good performance of the simple LP model, which assumes that work released in a planning period will be converted to output within the same period. This may seem to suggest that congestion is not particularly important in this experiment, but this is unlikely, since the average utilization in each period is in excess of 0.90, with considerable variation over time. However, under the given demand conditions the resource must operate close to its full capacity for most of the planning horizon, resulting in similar behavior for all planning models. Interestingly, all planning models underestimate the realized cost of the releases they propose. Detailed results are given in the original paper.

The presence of tool changes between different products changes the situation dramatically, as seen in Fig. 8.10. The performance of the ACF model collapses, which is not surprising since it was not designed to consider capacity losses of this type. Not only does it yield higher costs than other MDCFs, but the planning model severely underestimates the realized cost. The LP model uses a conservative estimate of capacity, based on the worst-case number of tool changes, resulting in poor cost performance but, interestingly, a very accurate prediction of the realized costs

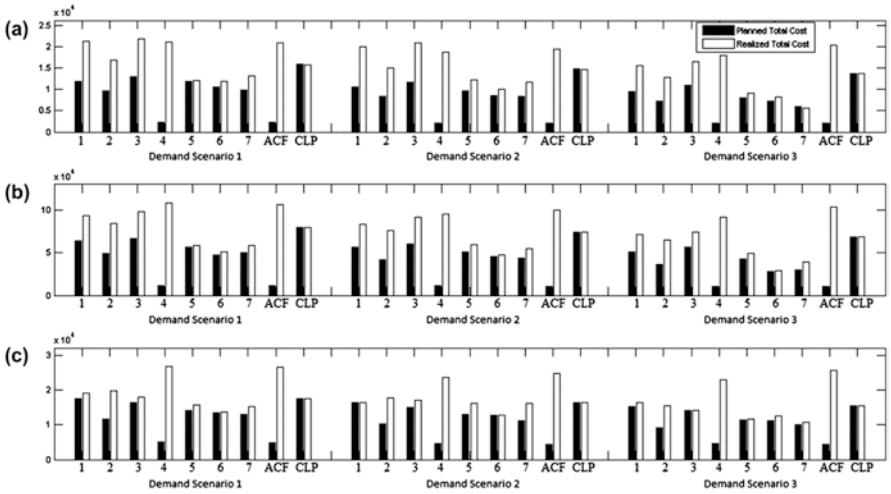

Fig. 8.10 Performance of various MDCFs relative to ACF and conservative LP Model. MDCFs 1 through 5 are O-based, MDCFs 6 and 7 are WIP-based

of the releases it generates. The W-MDCFs are now the best performers by a considerable margin, suggesting that in the presence of mix-dependent capacity losses, detailed representation of product mix is required. The poorer performance of the O-MDCFs is likely due to the fact that the output of a particular product depends on the amount of WIP of that product available during the period. The final experiments examine the impact of shop-floor dispatching with no tool changes. Similar to the findings of Asmundsson et al. (2006), the performance of the better MDCFs and ACF are generally comparable, suggesting that the use of non-delay dispatching policies in the absence of interference between products does not adversely affect the performance of ACF, while some of the MDCFs perform quite poorly. The poor performance of certain MDCFs is likely due to the release planning model converging to a poor local optimum rather than a global one.

The primary conclusion is that while MDCFs appear to be essential for good release planning in multiproduct systems where the processing time depends on the product mix, such as in the presence of setups, the resulting optimization models are substantially more involved than the linear programs resulting from the ACF model. The non-convex nature of these optimization models ought to come as no surprise to the reader; after all, even the univariate clearing functions discussed in Chap. 7 resulted in non-convex formulations in the presence of multiple products. The good news seems to be that for many functional forms, the non-convex behavior of the MDCFs seems rather benign, allowing locally optimal solutions obtained by conventional convex solvers to provide good performance. The development of efficient solution algorithms for these models, as quadratically constrained quadratic programs or bilinear models, remains an important topic for future work. The functional form of these MDCFs is also quite similar to those derived in the next chapter for lot-sizing problems.

In a subsequent paper, Albey et al. (2017) extend the idea of clearing functions from a single production resource to a production unit consisting of multiple resources, where in addition to requiring processing on several different resources, products also have routing flexibility that allows a given operation to be performed on one of several different machines. The objective of this work is to identify a set of state variables and a functional form for a MDCF that will allow the output of the overall production unit—not individual resources—in a planning period to be estimated to an acceptable degree of accuracy.

The point of departure for this work is the MDCF form (8.30), which was initially developed for a single production resource. In a production unit consisting of multiple resources, this functional form can be implemented at several levels of aggregation. The minimal unit of work is the machine-operation pair, specifying the processing of a particular operation of a specific product on a specific machine. In the presence of routing flexibility, a given operation may be performed on several alternative, non-identical machines. Operation-machine pairs can be summed for a specified operation, a specified machine, and over products. Summation over machines combines all operations processed on a given machine, while summation over a product sums the workload from all operations performed on that product. The reader will note we have met both these aggregations already: the single-variable clearing functions developed in Chap. 7 are based on aggregate workload over all operations processed at a given machine, while the product-based clearing functions of Kacar and Uzsoy (2014) aggregate the workload from all operations of a given product at a particular resource. The authors develop MDCFs for each of these levels of aggregation, and examine their performance in the presence of different levels of utilization and processing flexibility.

The release planning models based on the MDCFs follow the basic structure of other clearing function based models, with balance equations for finished goods inventory of each product and WIP of each of the basic units of aggregation. Thus in the model based on operation-machine pairs using the MDCF form (8.30), WIP balance equations are written for each operation at each machine as in the allocated clearing function model. When using the operation-based MDCF (8.29), WIP balance equations are written for each operation in each period. The P-MDCF requires WIP balance equations for each operation, since the WIP of each operation is weighted to reflect the likelihood of its emerging as finished product in the current period. As was the case for the single-stage systems, the resulting release planning models are non-convex, but are solved to a local optimum using the KNITRO convex nonlinear solver. The performance of the MDCFs is evaluated by the performance of the production unit under the releases developed by the release planning models using them. The univariate clearing function of Srinivasan et al. (1988) is used as a benchmark for comparison, and is implemented in a release planning model using the allocated clearing function formulation, but without piecewise linearization of the clearing functions. This will be referred to as the single-dimensional clearing function (SDCF) model in our discussions. The resulting nonlinear program is convex per the discussion in Chap. 7, and is solved to a global optimum by KNITRO. The production unit considered is a job shop consisting of six machines producing four products with different routings.

The first experiment examines the performance of the MDCFs as a function of utilization with no setups required between different products and no routing flexibility; each operation can be processed on exactly one machine. The findings from this experiment are confirmatory rather than surprising: at low to medium utilization levels, whose average across all machines and periods varies from 0.6 through 0.8, the performance of all MDCFs is fairly similar, with a slight advantage to OM-MDCF based on individual operation-machine pairs. The authors compare the planned and realized costs of the different models and find very close agreement for these utilization levels, indicating that the release planning model is able to accurately predict the consequences of its decisions on the shop floor. The SDCF also exhibits close agreement between planned and realized profit, but as average utilization reaches 0.8 it yields substantially lower profit than the MDCFs, suggesting once again that it systematically underestimates the capabilities of the production unit. This latter finding once again emphasizes the need for MDCFs when multiple products are present.

At higher average utilization levels, ranging from 0.9 through 1.1—the latter representing a major overload of the system—results are qualitatively different. Major differences appear between the different MDCFs. In terms of realized profit, OM-MDCF, the least aggregated of the MDCFs, is consistently the best performer. O-MDCF is the next best, followed by P-MDCF by a wide margin. The highly aggregated P-MDCF fails dramatically at these higher utilization levels, yielding extremely poor realized performance relative to the other MDCFs.

The second experiment in this study introduces routing flexibility by incrementally adding a single alternative machine for each operation of different products: first for the operations of Product 1, then Product 2, and then for all products. However, the choice of which of the alternative machines to use for a given operation is made by the shop-floor dispatching logic and is not available to the planning models. The improvement in performance of all models with the addition of even a limited amount of flexibility for a single product is quite striking, even when it affects only one of the four products. While the single-dimensional clearing function (SDCF) is the worst performer by a wide margin when there is no flexibility, the presence of flexibility for Product 1 alone more than doubles its expected profit. The marked improvement it obtained when flexibility is allowed for Product 1 suggests that most of its problems are due to the clearing functions estimated for the machines used by that product, machines 1, 3, and 5. The realized performances of OM-MDCF and O-MDCF are now very similar, and with the higher levels of flexibility even P-MDCF provides realized profit comparable to O-MDCF. This suggests that the presence of flexibility, in the form of alternative machines for specific operations of a product, allows capacity to be pooled across machines in a manner that makes it easier for the MDCFs, and even the SDCF, to predict.

The final experiment in this study examines the impact of setups between the different products, extending the analysis in Albey et al. (2014). Results for average utilization of 1.0 are shown in Fig. 8.11. Once again, under low average utilization all MDCFs and SDCF lead to quite similar performance, but as in the earlier study of single-stage systems the situation changes markedly at high utilization. As utilization increases, the more aggregated O-MDCF and P-MDCF begin to fall behind the less

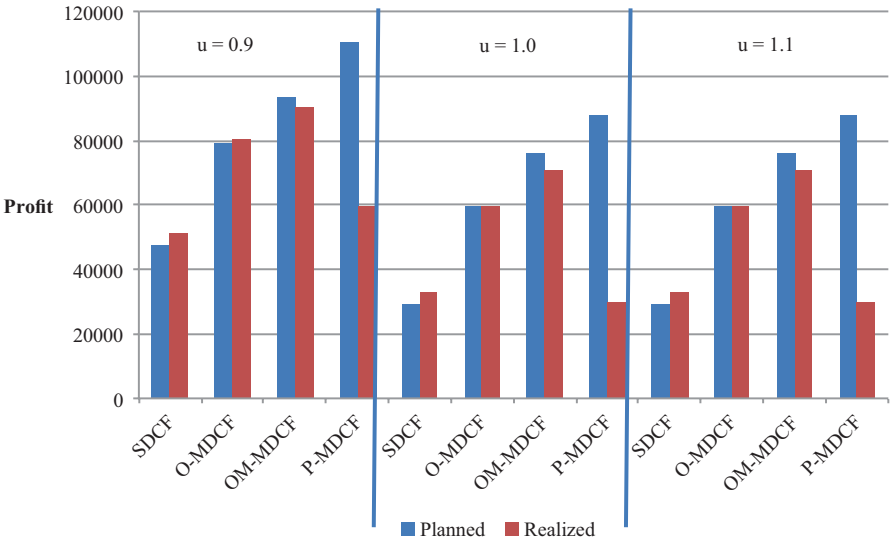

Fig. 8.11 Performance of MDCFS and SDCF with Setups under High Utilization

aggregated OM-MDCF in relative performance. The realized profit of all models decreases with increasing utilization, due to increasing backorder costs. The close agreement between the planned and realized costs of SDCF combined with its lower profit again suggests that it is underestimating the capabilities of the system, releasing less material which makes it easier for it to realize its planned profit, which remains substantially lower than those of the MDCFs. The very close agreement between the planned and realized profit of OM-MDCF and the very poor agreement for P-MDCF are equally interesting. Since setups are incurred by the processing of specific operations on specific machines, OM-MDCF is able to predict the potential output of each operation-machine pair quite accurately. P-MDCF fails since it does not capture operation-machine level data. O-MDCF occupies an intermediate position.

Once again, the performance of the MDCFs can be explained using intuition from queueing models. The impact of setups in a multiproduct system is to increase the variability of the effective processing time distribution, making it difficult for the more aggregated clearing functions to estimate the output of the system accurately over a wide range of product mixes and operating conditions. The reader will observe the recurring theme: the more different factors contributing to the variability of the effective processing time distribution at any resource, the harder for a clearing function with a few, aggregate state variables to estimate its output accurately. Setups are incurred on the basis of specific operations at specific machines, while machine failures affect all operations at a given machine in essentially the same way. One wishes that the authors had carried their experimental design to its logical conclusion, examining the impact of flexibility on problems with setups, and introducing machine failures. One would conjecture that if machine failures are the dominant source of variability, SDCF ought to perform fairly well, while if the primary source of variability is at the level of operation-machine pairs, OM-MDCF

ought to do better. The only situation in which P-MDCF might be expected to perform fairly well would be if all products required very similar processes in terms of both routing and operation processing times.

It is important to note that in this last experiment, as in the setup experiments reported for the single-stage systems, there is no attempt to perform any kind of lot sizing that would make use of setups with maximum efficiency; the cyclic sequence in which products are released approaches a worst-case situation in terms of the number of setups incurred. Clearly there is considerable scope for exploring the impact of different sequencing policies on the shape of the MDCF required to anticipate the behavior of the production unit.

8.5 Discussion

Our discussion of MDCFs has ranged over several different possibilities for extending the univariate clearing functions of Chap. 7, which have been the primary focus of research in this area for many years. Univariate clearing functions have been used to estimate the aggregate output of a production resource over all products, usually measured in units of time. For single product systems without sequence-dependent setup times this is, obviously, sufficient, but when multiple items compete for capacity additional logic is required. Chapter 7 discussed the difficulties encountered by univariate clearing functions in the presence of multiple products and presented the allocated clearing function formulation as an approximate, but generally effective, solution in the absence of setup times between products or under predetermined lot sizes. However, both queueing analysis and empirical observation suggest that when the amount of output the system can generate is significantly affected by the mix of the desired output, univariate clearing functions are inadequate.

The development of MDCFs requires additional state variables in the clearing functions, and the development of clearing functions estimating the output of each output item, which is usually a product but can also be defined as an operation-machine pair or multiple operation-machine pairs representing an operation that can be performed on alternative machines. The output capabilities of the system, whether a resource or a larger production unit, are captured by a system of MDCFs that jointly capture the tight interdependence of the output of different items. This approach requires the use of state variables related to each of the items produced, and use of state variables related to earlier planning periods has also been examined.

The common theme across experiments examining different MDCFs is that while univariate clearing functions are capable of estimating the aggregate output of a production resource or unit fairly accurately in the absence of setup times, their ability to estimate the mix of this output, at the level of individual items, is much more limited. Upon reflection, this should be no surprise; even in the absence of setups between products, the presence of multiple products with different service time distributions will increase the variability of the effective service time distribution, making it harder for a single-variable clearing function to produce accurate estimates of output under a wide range of operating conditions and product

mixes. In the presence of significant setup times, especially when lot sizes are determined at the scheduling level, the aggregate, univariate clearing functions fail dismally, as is only to be expected.

The use of MDCFs explicitly distinguishing between individual items yields more accurate output estimates, and hence better performance by the planning models that use them, but comes at the cost of significantly larger and more complex release planning models. In particular, the use of MDCFs results in non-convex optimization models that are significantly more difficult to solve than the linear programs of Chap. 5, or the convex nonlinear models and linear programs of Chap. 7. The nature of the non-convexity should be the subject of considerable future study. Anli et al. (2007) observe that non-convex behavior arises either when operating policies at the production units are "flagrantly suboptimal" or the items produced are highly diverse in nature, leading to a highly variable effective processing time distribution. Albey et al. (2014) also find that the objective function values obtained by the BARON global solver were the same as those from the KNITRO convex nonlinear solver in all cases where BARON converged to a solution. This further suggests that the non-convexity of these models is somewhat structured, raising the possibility that more efficient solution procedures may be possible. Further exploration of this issue is clearly an interesting direction for future research and provides a useful application of global optimization methods.

The development of clearing functions that explicitly recognize the transient state of the queues describing the system, without assuming steady-state behavior within the planning periods, has also raised a number of interesting issues. Both empirical evidence and queueing arguments demonstrate quite conclusively that the shape of the clearing function is different in the transient regime from the steady-state environment that is best studied. This issue of transient behavior is compounded by the fact that release planning models treat the planned state of the system in future periods as a deterministic parameter, while in reality these are better treated as (possibly biased) forecasts of random variables. The argument from Jensen's inequality suggests that even assuming unbiased forecasts of the future state variables, treating these estimates as deterministic parameters is likely to result in systematic overestimation of the output, an observation supported by considerable experimental evidence.

There also appears to be a basic tradeoff between the accuracy of the output predictions made by a clearing function and its computational tractability. In general, adding state variables and developing MDCFs for each item produced tend to improve the accuracy of the output predictions, but greatly increase the complexity of the resulting release planning models, as well as the complexity of fitting the MDCFs themselves. Advanced, high-dimensional machine learning techniques such as metamodeling of various kinds and neural networks may be able to produce quite accurate predictions of output, but are not amenable to incorporation in mathematical programming models of the kind this volume has focused on. The use of metamodels to accelerate simulation optimization approaches, by replacing a time-consuming simulation model with a fast running metamodel as in Li et al. (2016), suggests a possible way out of this dilemma, but considerable additional work is needed in this area.

References

Albey E, Bilge U, Uzsoy R (2014) An exploratory study of disaggregated clearing functions for multiple product single machine production environments. Int J Prod Res 52(18):5301–5322

Albey E, Norouzi A, Kempf KG, Uzsoy R (2015) Demand modeling with forecast evolution: an application to production planning. IEEE Trans Semicond Manuf 28(3):374–384

Albey E, Bilge U, Uzsoy R (2017) Multi-dimensional clearing functions for aggregate capacity modelling in multi-stage production systems. Int J Prod Res 55(14):4164–4179

Al-Khayyal FA (1992) Generalized bilinear programming: part I. Models, applications and linear programming relaxation. Eur J Oper Res 60(3):306–314

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

Ankenman BE, Bekki JM, Fowler J, Mackulak GT, Nelson BL, Yang F (2010) Simulation in production planning: an overview with emphasis in recent developments in cycle time estimation. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning production and inventories in the extended enterprise: a state of the art handbook, vol 1. Springer, New York, pp 565–592

Anli OM, Caramanis M, Paschalidis IC (2007) Tractable supply chain production planning modeling non-linear lead time and quality of service constraints. J Manuf Syst 26(2):116–134

Aouam T, Uzsoy R (2012) Chance-constraint-based heuristics for production planning in the face of stochastic demand and workload-dependent lead times. In: Armbruster D, Kempf KG (eds) Decision policies for production networks. Springer, London, pp 173–208

Aouam T, Uzsoy R (2015) Zero-order production planning models with stochastic demand and workload-dependent lead times. Int J Prod Res 53(6):1–19

Armbruster D, Fonteijn J, Wienke M (2012) Modeling production planning and transient clearing functions. Logistics Res 5:133–139

Askin RG, Hanumantha GJ (2018) Queueing network models for analysis of nonstationary manufacturing systems. Int J Prod Res 56(1-2):22–42

Askin RG, Standridge CR (1993) Modeling and analysis of manufacturing systems. Wiley, New York

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

Bao X, Sahinidis NV, Tawarmalani M (2011) Semidefinite relaxations for quadratically constrained quadratic programming: a review and comparisons. Math Program Ser B 129:129–157

Billingsley B (1995) Probability and measure. Wiley, New York

Bischoff W (2017) Numerical tests of order release models with one- and two-dimensional clearing functions. Department of Information systems, production and logistics management. University of Innsbruck, Innsbruck, Austria

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

Cohen JW (1969) The single server queue. North-Holland, Amsterdam

Fine CH, Graves SC (1989) A tactical planning model for manufacturing subcomponents of mainframe computers. J Manuf Oper Manag 2:4–34

Hadley G (1964) Nonlinear and dynamic programming. Addison-Wesley, Reading, MA

Häussler S, Missbauer H (2014) Empirical validation of meta-models of work centres in order release planning. Int J Prod Econ 149:102–116

Heath DC, Jackson PL (1994) Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. IIE Trans 26(3):17–30

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

Kacar NB, Uzsoy R (2014) A comparison of multiple linear regression approaches for fitting clearing functions to empirical data. Int J Prod Res 52(11):3164–3184
Kacar NB, Uzsoy R (2015) Estimating clearing functions for production resources using simulation optimization. IEEE Trans Autom Sci Eng 12(2):539–552
Kacar NB, Irdem DF, Uzsoy R (2012) An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms. IEEE Trans Semicond Manuf 25(1):104–117
Kacar NB, Moench L, Uzsoy R (2013) Planning wafer starts using nonlinear clearing functions: a large-scale experiment. IEEE Trans Semicond Manuf 26(4):602–612
Kacar NB, Moench L, Uzsoy R (2016) Modelling cycle times in production planning models for wafer fabrication. IEEE Trans Semicond Manuf 29(2):153–167
Karmarkar US (1989) Capacity loading and release planning with work-in-progress (WIP) and lead-times. J Manuf Oper Manag 2(1):105–123
Li M, Yang F, Uzsoy R, Xu J (2016) A metamodel-based monte carlo simulation approach for responsive production planning of manufacturing systems. J Manuf Syst 38:114–133
Lin PC, Uzsoy R (2016) Chance-constrained formulations in rolling horizon production planning: an experimental study. Int J Prod Res 54(13):3927–3942
Linderoth J (2005) A simplicial branch-and-bound algorithm for solving quadratically constrained quadratic programs. Math Program Ser B 103:251–282
Missbauer H (1998) Bestandsregelung als Basis für eine Neugestaltung von PPS-Systemen. Physica, Heidelberg
Missbauer H (2007) Durchlaufzeitmanagement in Dezentralen PPS-Systemen. In: Corsten H, Missbauer H (eds) Produktions- und Logistikmanagement. Festschrift für Günther Zäpfel zum 65. Geburtstag. Verlag Franz Vahlen GmbH, Munich
Missbauer H (2009) Models of the transient behaviour of production units to optimize the aggregate material flow. Int J Prod Econ 118(2):387–397
Missbauer H (2011) Order release planning with clearing functions: a queueing-theoretical analysis of the clearing function concept. Int J Prod Econ 131(1):399–406
Missbauer H (2014) From cost-oriented input-output control to stochastic programming? Some reflections on the future development of order release planning models. In: Gössinger R, Zäpfel G (eds) Management Integrativer Leistungserstellung. Festschrift Für Hans Corsten. Duncker & Humblot GmbH, Berlin, pp 525–544
Missbauer H, Stolletz R (2016) Order release optimization for time-dependent and stochastic manufacturing systems. University of Innsbruck and University of Mannheim. 26 pp
Orcun S, Uzsoy R (2011) The effects of production planning on the dynamic behavior of a simple supply chain: an experimental study. In: Kempf KG, Keskinocak P, Uzsoy R (eds) Planning in the extended enterprise: a state of the art handbook. Springer, Berlin, pp 43–80
Papadopoulos HT, Heavey C, Browne J (1993) Queueing theory in manufacturing systems analysis and design. Chapman & Hall, London [u.a.]
Selçuk B, Fransoo JC, De Kok AG (2008) Work-in-process clearing in supply chain operations planning. IIE Trans 40(3):206–220
Shingo S (1986) A revolution in manufacturing: the SMED system. Productivity Press, Cambridge
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
Stolletz R (2008) Approximation of the non-stationary M(t)/M(t)/c(t)-queue using stationary queueing models: the stationary backlog-carryover approach. Eur J Oper Res 190(2):478–493
Stolletz R, Lagershausen S (2013) Time-dependent performance evaluation for loss-waiting queues with arbitrary distributions. Int J Prod Res 51(5):1366–1378
Tawarmalani M, Sahinidis NV (2005) A polyhedral branch and cut approach to global optimization. Math Program 103(2):225–249
Yang F, Ankenman B, Nelson BL (2006) Efficient generation of cycle time-throughput curves through simulation and metamodeling. Naval Res Logistics 54(1):78–93