Chapter 9
Lot-Sizing Models Using Multi-dimensional Clearing Functions

The order release models described in this volume rely heavily on the functional relationship between the expected output of a production resource and its expected workload which, as discussed in Chap. 2 for the case of steady-state queues, is related to the expected cycle time by Little's Law. This relationship is significantly affected by various decision rules used within the PPC system, such as scheduling policies on the shop floor. *Lot sizing*, the decision as to how much of a product to produce each time a machine is set up for the product, is of particular importance in this respect. For a given production quantity, determined by the master production schedule, the lot sizes influence capacity utilization (via the amount of setup time required on the resource in a planning period), the mean and variance of the interarrival times (via the number and size of production lots), and the mean and variance of the service times (via the lot sizes). Lot-sizing models were among the earliest mathematical formulations of production planning problems (Harris 1915). The extensive literature on deterministic lot-sizing problems (Drexl and Kimms 1997; Brahimi et al. 2006; Pochet and Wolsey 2006; Quadt and Kuhn 2008) has generally focused on the tradeoff between fixed setup or ordering costs and inventory holding costs without considering the effects of congestion. The relationship between the lot size Q and average cycle time has been explored from several angles, including simultaneous lot sizing and scheduling (Drexl and Kimms 1997) and lot streaming (Missbauer 2002; Jen Huei and Huan Neng 2005; Cheng et al. 2013). Following the discussion in Chap. 2, we begin this section with insights from simple queueing models, and then show how these can be used to develop a system of multivariate clearing functions to address a dynamic lot-sizing problem.

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_9

9.1 Impact of Lot Sizes on the Performance of Production Resources

As the extensive literature on deterministic economic order quantities would suggest, lot-sizing decisions have significant impact on the behavior of production systems even in completely deterministic environments. Karmarkar (1989) proposes an example by considering a synchronous production line with N stations producing a single item in batches of a fixed size of Q units. Batches are transferred to the next station at the completion of processing, and a setup time of S time units is required for each batch. The production rate at each station is assumed to be P units/time unit. Thus each batch has a cycle time of $(S + Q/P)$ time units at each station, and a total cycle time in the line of $T = N(S + Q/P)$ time units. Since under synchronous operation there will be no queueing, and one batch will complete its processing and leave the system every T/N time units, the average output rate of the line will be

$$X = \frac{Q}{\left(S + \dfrac{Q}{P}\right)} = \frac{PQ}{PS + Q} \tag{9.1}$$

which is a saturating, non-decreasing function of the lot size. The non-decreasing, saturating nature of the function arises from the fact that as the lot size Q increases, the setup time per part, which constitutes a loss of production capacity, is reduced, eventually reaching 0 as $Q \to \infty$.

Karmarkar (1987) and Zipkin (1986) were among the first to study the relationship between lot size and mean flow time in an *M*/*M*/*1* queueing system operating at a fixed output rate. Karmarkar (1987) derived this relationship for a single-server queue producing a single item. We follow this derivation for an *M*/*G*/*1* system using the following notation:

D: total demand per period (in product units)
p: processing time per unit
S: setup time per lot
λ: arrival rate of the lots at the server
t_e, σ, c_e: mean, standard deviation, and coefficient of variation, respectively, of the service times of the lots, given by the sum of setup and processing times
Q: lot size, assumed to be identical for all lots

For brevity of exposition, we will assume the coefficient of variation c_e of the service times is independent of the lot size Q. Then the expected arrival rate of lots at the machine is given by $\lambda = D/Q$ and the expected service time of a lot $t_e = S + pQ$, yielding a utilization of $u = \lambda t_e = D(p + S/Q)$. Assuming a Poisson arrival process, the Pollaczek–Khintchine formula (Medhi 1991) gives the mean queue (waiting) time of a lot as

$$T_Q = \frac{u^2 + \lambda^2\sigma^2}{2\lambda(1-u)} = \frac{D\left(c_e^2+1\right)\left(pQ+S\right)^2}{2\left[Q\left(1-pD\right)-DS\right]} \tag{9.2}$$

and the mean cycle time as

$$T = T_Q + S + pQ \tag{9.3}$$

Both the mean queue time (9.2) and the mean cycle time (9.3) are convex functions of the common lot size Q. When different lot sizes Q_j for multiple products j are used, the mean waiting time remains a convex function of the lot sizes, but the mean cycle time is non-convex (Karmarkar et al. 1992).

The single-product cycle time (9.2) illustrates an intuitive phenomenon termed the "Process Batching Law" by Hopp and Spearman (2008): the presence of positive setup times imposes a lower limit on the lot sizes, and as the lot size approaches this limit from above the utilization $u \to 1$, and hence the average cycle time $T \to \infty$. If the lot size becomes large, i.e., $Q \to \infty$, the impact of the setup time vanishes and T increases asymptotically proportionally with the lot size. As a consequence of this structure, Karmarkar (1987) shows that T is minimized for a well-defined lot size.

These insights have been refined and extended in subsequent work that is beyond the scope of this volume (Wijngaard 1989; Benjaafar 1996; Missbauer 2002; Jutz 2017). Extensions include considering multiple products, more complex material flow structures, and its integration with the inventory control system that determines the arrival process of the lots (Zipkin 1986; Vaughan 2006). The modeling approach can also be extended to multistage systems with stage-specific lot sizes (Missbauer and Jutz 2018).

In order to develop clearing function models, this model must be reformulated to express the expected output as a function of lot sizes and expected WIP. By the PASTA (Poisson Arrivals See Time Averages) property of the arrival process (Buzacott and Shanthikumar 1993, p. 54), the average *actual* waiting time of the customers (lots) T_Q is identical to the average *virtual* waiting time at time t, defined as the waiting time that would be seen by a customer arriving at time t. For a single-server system, the average virtual waiting time is identical to the average WIP at the server, measured in hours of work (average remaining work). Using (9.2) the expected output pD, excluding time spent in setups and expressed in hours of work, can be written as:

$$X = pD = \frac{2pQT_Q}{\left(pQ+S\right)\left[2T_Q + \left(c_e^2+1\right)\left(pQ+S\right)\right]} \tag{9.4}$$

Equation (9.4) implies that higher service time variability reduces the output for a given average WIP. The impact of the lot size Q on the relationship between average WIP and output is shown in Fig. 9.1 for different lot sizes.

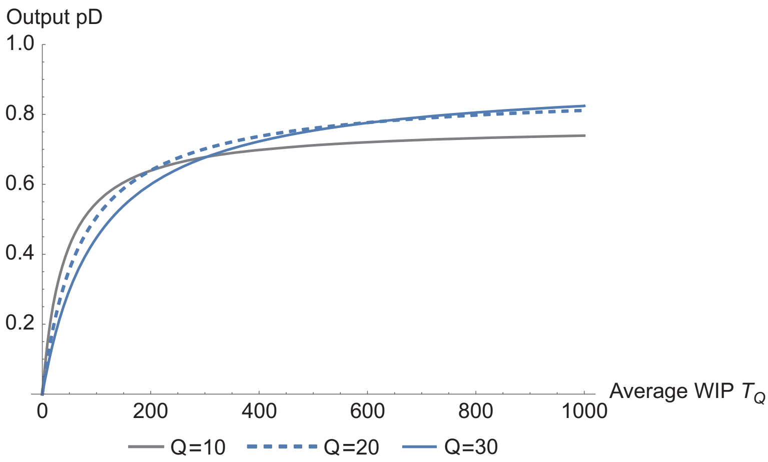

Fig. 9.1 Expected output as a function of expected WIP for different values of the lot size Q ($p = 5, S = 15, c_e = 0.5$)

This modeling approach can be applied in two ways. The relationship between lot sizes and average flow time can be used to derive standard lot sizes that yield a good compromise between the potentially conflicting goals of reducing cycle times on the one hand and minimizing setup and cycle inventory holding costs on the other (Missbauer 2002). The actual lot sizes implemented on the shop floor can be determined by modifying these standard lot sizes based on short-term demand information, leading to a hierarchical lot-sizing system (Söhner and Schneeweiss 1995). The benefit of modifying the standard lot sizes has been questioned in the literature (Wijngaard 1989). Within this decision structure the lots to release are determined outside the release model and consume the release quantities R_{jt} calculated by the release model as described in Chaps. 5 through 8.

An alternative approach is to determine lot sizes and order releases simultaneously using a release model with a multi-dimensional clearing function that includes some measure of workload and the lot sizes as state variables determining the expected output in the spirit of (9.4). We discuss a model of this type in the next section.

9.2 A MDCF Model for Lot Sizing

In this section, we present a single-stage multi-item dynamic lot-sizing model developed by Kang et al. (2014) where the production resource is modeled as an *M*/*G*/*1* queue. The behavior of the system is modeled by a set of multi-dimensional clearing functions (MDCFs) derived by steady-state queueing analysis, instead of the empirically estimated MDCFs described in the previous chapter.

We consider a single production resource processing N different products $i = 1,\ldots, N$ with deterministic processing time p_i and sequence-independent setup time s_i that is incurred whenever a unit of product i is processed after completion of a different product. The planning horizon is divided into T discrete time periods of uniform length, and all processing and setup times are expressed in units of this planning period length. Lots of product i arrive at the resource following a Poisson process with rate λ_i. Due to the random arrival process, the service time is a random variable. In order to address the lot-sizing problem, the MDCFs describing the output of the resource must reflect the lot sizes. This is accomplished by assuming that the planning periods are sufficiently long that the system is in steady state, and following the analysis of Karmarkar (1987) and Karmarkar et al. (1992). Since we derive the MDCFs for a generic planning period, the period index is dropped in the following analysis.

The deterministic processing time of a lot of Q_i units of product i is given by:

$$P_i = s_i + p_i Q_i \tag{9.5}$$

Since lots of product i arrive following a Poisson process with rate λ_i, the probability that a randomly selected batch is of product i is given by λ_i/λ, where

$$\lambda = \sum_{i=1}^{N} \lambda_i \tag{9.6}$$

Thus the mean and variance of the random variable P denoting the processing time of lots at the resource are given by:

$$E\left[P\right] = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} P_i \quad \text{and} \quad E\left[P^2\right] = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} P_i^2 \tag{9.7}$$

It is a standard result in queueing theory (Buzacott and Shanthikumar 1993, p. 62) that the expected waiting time for the *M/G/1* queue is given by:

$$T_Q = \frac{\lambda E\left[P^2\right]}{2\left(1-u\right)} = \frac{\lambda \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} P_i^2}{2\left(1-u\right)} = \frac{\sum_{i=1}^{N} \lambda_i P_i^2}{2\left(1-u\right)} \tag{9.8}$$

where u denotes the average utilization as in previous chapters. The expected cycle time of product i is then given by $\tau_i = P_i + T_Q$. Little's Law then yields

$$\lambda_i = \frac{\left(\frac{\bar{W}_i}{Q_i}\right)}{T_Q + P_i} \tag{9.9}$$

where $\bar{W}_i$ denotes the time-average WIP of product i over the duration of the planning period. Since we assume the system is in steady state, the number Y_i of lots of product i produced during the period can be substituted for λ_i, yielding

$$Y_i = \frac{\left(\frac{\bar{W}_i}{Q_i}\right)}{\tau_i} = \frac{\left(\frac{\bar{W}_i}{Q_i}\right)}{P_i + T_Q} = \frac{\left(\frac{\bar{W}_i}{Q_i}\right)}{P_i + \frac{\sum_{i=1}^{N} Y_i P_i^2}{2(1-u)}} = \frac{\left(\frac{\bar{W}_i}{Q_i}\right)}{P_i + \frac{\sum_{i=1}^{N} Y_i P_i^2}{2\left(1-\sum_{i=1}^{N} Y_i P_i\right)}} \tag{9.10}$$

Noting that all processing times are in units of the planning period, and multiplying both sides of (9.10) by Q_i, we obtain the total number of units of product i produced in the planning period as

$$f_i\left(Q_i, \bar{W}_i, Y_i, Q_i', W_i', Y_i'\right) = Q_i Y_i = \frac{\bar{W}_i}{P_i + \frac{u_i P_i + \sum_{j \neq i} u_j P_j}{2\left(1 - u_i - \sum_{j \neq i} u_j\right)}} \tag{9.11}$$

which can be written out as

$$f_i\left(Q_i, \bar{W}_i, Y_i, Q_i', W_i', Y_i'\right) = Q_i Y_i$$
$$= \frac{\bar{W}_i}{\left(s_i + p_i Q_i\right) + \frac{Y_i\left(s_i + p_i Q_i\right)^2 + \sum_{j \neq i} Y_j\left(s_j + p_j Q_j\right)^2}{2\left(1 - \left(s_i + p_i Q_i\right) - \sum_{j \neq i} Y_j\left(s_j + p_j Q_j\right)\right)}} \tag{9.12}$$

where Q_i' denotes the vector of lot sizes Q_j for all products except i, and $\overline{W'}_i$ and Y_i' are defined analogously. The MDCF (9.12) is an ugly, non-convex expression, but is actually quite intuitive: The output of a particular product i in a planning period depends on its own lot size Q_i, the number of lots produced Y_i, and its time-average WIP level $\bar{W}_i$, as would be expected in a single-product model. However, it is also affected by the lot sizes, WIP levels, and number of lots of all other products. As seen in the intermediate expression (9.11), this is because these quantities determine the fraction of current machine utilization available to the product i in the planning period. Thus the output mix of the machine is jointly determined by the set of N MDCFs (9.12). The explicit consideration of lot sizing has resulted in the addition of state variables reflecting the lot sizes of each product during the planning period. An example of this MDCF is illustrated in Fig. 9.2; note that the level sets shown on the horizontal plane, which are the feasible combinations of WIP and lot sizes that yield the specified output, match those given by Karmarkar (1987).

Anli et al. (2007) present a MDCF with similar state variables, but take a very different approach to estimating it; they use an iterative approach between the

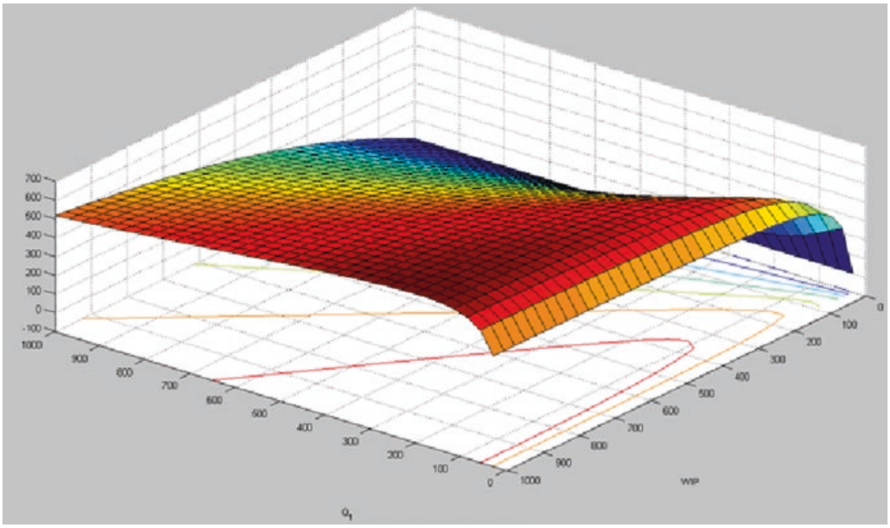

Fig. 9.2 Illustration of MDCF with lot sizing for two-product system: output of product 1 as a function of its lot size and WIP for fixed lot size and output of product 2 (Kang et al. 2014)

individual production units for which the MDCF is being developed and the goods flow model. Tentative release plans are computed by the planning level, which are then used by the production units to estimate their realized performance. These realized performance estimates are then fed back to the planning level, which generates additional constraints derived from these estimates to refine its models of the capabilities of the production units. In the language of Schneeweiss (2003), tentative release plans are communicated to the production units and their feedback is then used to refine the planning level's anticipation functions for the production units. The approach of Anli et al. (2007) is unique in presenting an integrated, well thought out decomposition of the supply chain planning problem into multiple subproblems, including the goods flow problem, safety stock levels, and MDCFs for the individual production units, with promising computational results.

Several aspects of this MDCF are worthy of comment. Like Karmarkar (1987), it highlights the strong interdependence of products in a multiproduct queueing system: decisions made for any product, such as the level of output or the lot size, affect all other products. The use of this MDCF for a planning period of fixed finite length is clearly heuristic; the derivation assumes the queue is in steady state during the planning period, which is unlikely to be the case in general. The model also assumes that the lot sizes are decision variables associated with each planning period, and hence that these can be changed by management in each planning period. This is clearly possible for newly released orders, but it is unlikely that lots already released to production can be reconfigured without considerable disruption of ongoing operations. If the cycle time of some fraction of lots in each period exceeds the length of a planning period, it is thus likely that there will be lots of different sizes on the shop floor at least some of the time; the transient state refers not only to the number of orders at the workcenters but to the composition of the order sizes as well.

The MDCFs (9.12) can be incorporated into an integrated release planning and lot-sizing model in a straightforward manner, using the following notation:

Decision variables:
Y_{it}: number of lots of product i produced in period t
Q_{it}: lot size of product i in period t
I_{it}: finished goods inventory of product i at the end of period t
B_{it}: amount of product i backlogged at the end of period t
W_{it}: WIP of product i at the end of period t
$\bar{W}_{it}$: time-average WIP level of product i during period t
R_{it}: number of units of product i released in period t

Parameters:
h_{it}: unit finished goods holding cost in period t
w_{it}: unit WIP holding cost in period t
b_{it}: unit backlogging cost in period t

The model can then be written as follows:

$$\min \sum_{i=1}^{N}\sum_{t=1}^{T}\left[h_{it}I_{it} + b_{it}B_{it} + w_{it}W_{it}\right] \tag{9.13}$$

subject to

$$I_{it} - B_{it} = I_{i,t-1} - B_{i,t-1} + Q_{it}Y_{it} - D_{it}, \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.14}$$

$$W_{it} = W_{i,t-1} + R_{it} - Q_{it}Y_{it}, \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.15}$$

$$\bar{W}_{it} = \frac{W_{it} + W_{i,t-1}}{2}, \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.16}$$

$$Q_{it}Y_{it} \le f_i\left(Q_t, \bar{W}_t, Y_t\right), \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.17}$$

$$\sum_{i=1}^{N} Y_{it}\left(s_i + p_i Q_{it}\right) < 1, \quad t = 1,\ldots,T \tag{9.18}$$

$$Q_{it}, Y_{it}, R_{it}, I_{it}, W_{it}, B_{it} \ge 0, \quad \text{integer} \tag{9.19}$$

Constraints (9.18) are a stability condition that is redundant when the MDCFs (9.17) are present; it is included in the model to help reduce the solution time. The model (9.13)–(9.19) is a single-stage multi-item dynamic lot-sizing model, with some interesting differences. The presence of the MDCFs leads to non-convex constraints, even when the integrality constraints are relaxed. In addition, traditional lot-sizing models focus on the tradeoff between the fixed cost of setups and inventory holding costs, while in this model setup costs are conspicuous by their absence. It can be argued that the actual cash costs of setup changes are relatively small and are usually limited to the scrap generated while adjusting the machine and tooling to the new product. In the short term, labor and machinery are all fixed costs, so the main component of a setup cost in a production environment is the opportunity cost of the lost production time. This opportunity cost, however, is difficult to estimate in practice. If the facility has sufficient excess capacity that the setup will not result in any loss of revenue, the opportunity cost of capacity associated with the setup is clearly zero; this is equally clearly not the case if the facility is highly utilized and setups result in lost sales due to reduced output.

Due to the complexity of the integer nonlinear program (9.13)–(9.19), Kang et al. (2014) relax the integrality constraints, solve the resulting non-convex model to a local optimum and then heuristically round the resulting fractional solution to an integer feasible solution. In a later paper (Kang et al. 2018), they propose a more sophisticated rounding heuristic that gives considerably improved solutions over the original approach. Due to the absence of setup costs, the performance of the model is compared to that of a model due to Erenguc and Mercan (1990), which requires some additional notation:

Decision variables:

K_{it}: binary variable equal to 1 if a setup is performed for product i in period t, and zero otherwise

X_{it}: amount of product i produced in period t

Parameters:

M: a very large number

The model can be stated as follows:

$$\min \sum_{i=1}^{N}\sum_{t=1}^{T}\left[h_{it}I_{it} + b_{it}B_{it}\right] \tag{9.20}$$

subject to:

$$I_{it} = I_{i,t-1} + X_{it} - D_{it}, \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.21}$$

$$X_{it} \le MK_{it}, \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.22}$$

$$\sum_{i=1}^{N}\left[K_{it}s_{it} + p_iQ_{it}\right] \le 1, \quad t = 1,\ldots,T \tag{9.23}$$

$$K_{it} \in \{0,1\}, I_{it}, X_{it} \ge 0, \quad i = 1,\ldots,N,\ t = 1,\ldots,T \tag{9.24}$$

There are some interesting contrasts between this model and the MDCF-based model (9.13)–(9.19). The model of Mercan and Erenguc assumes that all production of a given product in a given period will be processed as a single lot, while the MDCF-based model allows multiple smaller lots. The Mercan–Erenguc model does not consider queueing effects at all, while these are central to the MDCF-based model. In fairness, the Erenguc–Mercan model was never intended to be used in a queueing environment, but rather for big-bucket lot-sizing or product cycling problems where queueing does not arise.

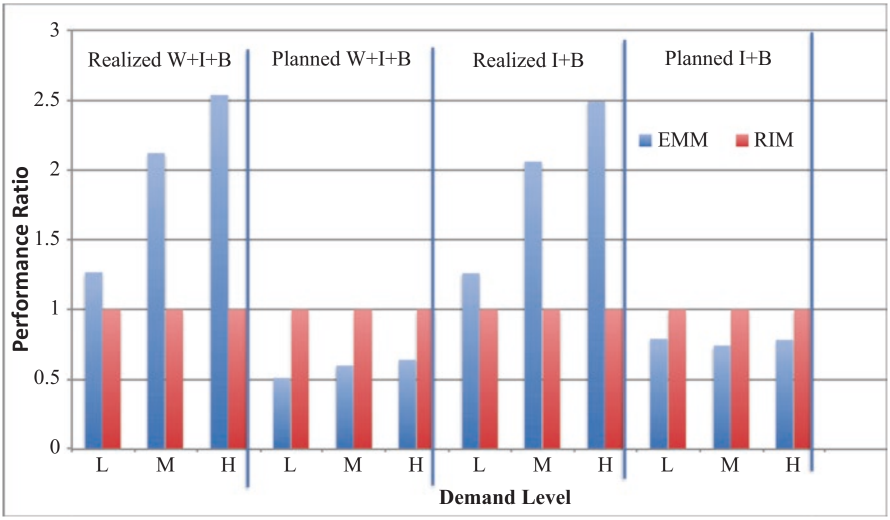

Fig. 9.3 Performance comparison of Erenguc–Mercan model (EMM) and MDCF-based lot-sizing model (RIM)

The logical way to evaluate the performance of this integrated release planning and lot-sizing model is to simulate the behavior of the production system operating under the lot sizes and release quantities it suggests. Details of the computational experiments are given in Kang et al. (2014), but representative findings are summarized in Fig. 9.3. The planned quantities refer to the objective function values from the mathematical models, while the realized values are those observed when the decisions from the mathematical models are implemented. Since the Erenguc–Mercan model does not consider congestion, and hence ignores WIP, we report the objective functions with and without WIP costs to observe how well the mathematical models predict the consequences of their decisions. The simulation model relaxes the assumption of a constant lot size in each period; if all lots released in a given period have not exited the system by the start of the following period, lots with different sizes will coexist in the system.

It is clear from the figure that the failure of the Erenguc–Mercan model to consider WIP results in the MDCF model performing considerably better. The planned objective function of the Erenguc–Mercan model, which considers only inventory and backorder costs, is actually quite close to those components of the planned cost from the MDCF model. However, the ability of the MDCF model to produce the demand for a given period in a number of small batches results in considerable improvement in cycle times, and major differences in performance between the two models. Although it is not evident from the limited data shown, the differences between the two models are largest at low to medium demand levels. At high demand, and hence utilization, lot sizes have to be large in order for the system to meet demand. Hence all production of a product in a given period is processed in a single lot, as required by the Erenguc–Mercan model. At lower utilization levels, however, the MDCF model can take advantage of the available excess capacity by using smaller batches with more setups, resulting in lower flow times and better performance.

As discussed above, the development of this model rests on a number of heuristic assumptions: the use of steady-state queueing models to derive the MDCF and the approximate solution of the resulting nonlinear integer program by solving its continuous relaxation and rounding to an integer feasible solution. There is no doubt that each of these introduces errors, which are likely to grow as the length of the planning period decreases. However, the MDCF model is in any case unsuitable for short-term release planning due to the difficulty of adjusting lot sizes on the shop floor after release. The model is better viewed as a longer-term aggregate model that can be used to examine the impact of lot sizes in the presence of changing demand conditions. It is also likely that some of the more egregious errors introduced by these assumptions are remedied to some degree by the myopic rounding scheme implemented at the execution level in the simulation model.

9.3 Insights from a MDCF-Based Lot-Sizing Model

The MDCF-based lot-sizing model (9.13)–(9.19) is clearly an extension to the multi-item capacitated lot-sizing problem of the type studied by Billington et al. (1983) and Trigeiro et al. (1989) and reviewed extensively by Quadt and Kuhn (2008). These models, along with their many successors, focus on the tradeoff between the fixed costs of setups and inventory holding costs, while considering capacity constraints without congestion as reflected by constraint (9.23) in the Erenguc–Mercan model above. In this section, we present a column generation heuristic for the MDCF-based lot-sizing model developed by Kang et al. (2011), with the purpose of providing insight into the practical difficulties of estimating setup costs in production environments. Similar column generation approaches for capacitated lot-sizing problems without congestion have been developed by Lasdon and Terjung (1971) and de Graeve and Jans (2007).

The basic idea of column generation approaches for capacitated lot-sizing problems is to decompose the problem into a master problem that allocates capacity among the N different products, and pricing subproblems that perform the optimal lot sizing for each product subject to the capacity allocation given by the master problem. Hence, in the Lasdon–Terjung approach (Lasdon and Terjung 1971), when the master problem allocates capacity, the pricing subproblems are single-item uncapacitated dynamic lot-sizing problems whose objective function is modified by the dual prices obtained from the master problem. Detailed presentations of Dantzig–Wolfe decomposition and column generation methods can be found in Desaulniers et al. (2005) and Lasdon (1970).

Following the usual approach to developing a column generation approach, let us denote the set of all feasible schedules for product i, $i = 1, \ldots, N$, by Υ_i. Since all decision variables associated with a product i in the model (9.13)–(9.19) must take integer values, the sets Υ_i, $i = 1, \ldots, N$ will each consist of a very large number of discrete schedules. Let τ_i^k denote a column vector with T entries associated with a solution $k \in \Upsilon_i$ whose tth entry is the capacity required by product i in period t for schedule k, given by

$$\tau_{it}^k = \left(s_i + p_i Q_{it}^k\right) Y_{it}^k \tag{9.25}$$

where Q_{it}^k denotes the lot size of product i in period t in the schedule $k \in \Upsilon_i$ and Y_{it}^k denotes the number of lots of product i produced in period t in this schedule. We also define the cost vector V_i^k as a column vector with T entries.

$$V_{it}^k = h_{it} I_{it}^k + w_{it} W_{it}^k + b_{it} B_{it}^k \tag{9.26}$$

Defining the decision variables

$$\gamma_k^i = \begin{cases} 1, \text{ if schedule } k \in \Upsilon_i \text{ is selected for product } i \\ 0, \text{ otherwise} \end{cases} \tag{9.27}$$

we can rewrite the model (9.13)–(9.19) as that of selecting exactly one schedule for each product such that the resulting schedules are capacity feasible and the objective function is minimized. The resulting master problem is given by:

(Master Problem: MP)

$$\min \sum_{i=1}^{N} \sum_{k \in \Upsilon_i} V_i^k \gamma_i^k \tag{9.28}$$

subject to

$$\sum_{k \in \Upsilon_i} \tau_{it}^k \gamma_i^k \le C_t, \quad i = 1, \ldots, N,\ t = 1, \ldots, T \tag{9.29}$$

$$\sum_{k \in \Upsilon_i} \gamma_i^k = 1, \quad i = 1, \ldots, N \tag{9.30}$$

$$\gamma_i^k \in \{0,1\}, \quad k \in \Upsilon_i,\ i = 1, \ldots, N \tag{9.31}$$

Since this is a binary set covering problem that is hard to solve, we relax the integrality constraints (9.31), replacing them with

$$0 \le \gamma_i^k \le 1, \quad i = 1, \ldots, N,\ k \in \Upsilon_i \tag{9.32}$$

to obtain the relaxed master problem (RMP).

$$\min \sum_{i=1}^{n} \sum_{k \in \Upsilon_i} V_i^k \gamma_i^k \tag{9.33}$$

subject to

$$\sum_{i=1}^{N}\sum_{k\in\Upsilon_i}\tau_{it}^{k}\gamma_i^{k}\le C_t,\quad t=1,\ldots,T \tag{9.34}$$

$$\sum_{k\in\Upsilon_i}\gamma_i^{k}=1,\quad i=1,\ldots,N \tag{9.35}$$

$$0\le\gamma_i^{k}\le 1,\quad i=1,\ldots,N,\ k\in\Upsilon_i \tag{9.36}$$

Since enumerating all columns in the RMP is impractical, we use a restricted relaxed master problem (RRMP) with a limited number of columns that are generated by a column generation approach. The RRMP is initialized with an initial set of columns and solved to optimality. However, this solution is only optimal with respect to the limited set of columns considered in the RRMP; there may yet exist columns in some of the sets Υ_i that have not yet entered the RRMP, but which might improve the objective function if they were to enter, i.e., have negative reduced costs. A pricing subproblem is thus solved for each product $i = 1,\ldots, N$ to determine whether any columns with negative reduced costs exist.

To formulate the pricing subproblem for product i, we define α_{it}^{k} to be the dual variable associated with the capacity constraints (9.34) and $\propto_i^{k}$ those associated with constraints (9.35). Then the reduced cost for a new column to enter the basis of the RRMP will be

$$\sum_{t=1}^{T}\left[V_{it}^{k}+\alpha_{it}^{k}\left(s_i+p_iQ_{it}^{k}\right)Y_{it}^{k}\right]+\mu_i^{k} \tag{9.37}$$

The pricing subproblem seeks a schedule k for product i such that the reduced cost is negative; if no such schedule can be found for any product an optimal solution to the relaxed master problem has been obtained. We can thus state the pricing subproblem for product i, $i = 1,\ldots, N$, as follows:

$$\min\ \sum_{t=1}^{T}\left[V_{it}^{k}+\alpha_{it}^{k}\left(s_i+p_iQ_{it}^{k}\right)Y_{it}^{k}\right]+\mu_i^{k} \tag{9.38}$$

subject to

$$W_{it}^{k}=W_{i,t-1}^{k}+R_{it}^{k}-Q_{it}^{k}Y_{it}^{k},\quad i=1,\ldots,N,\ t=1,\ldots,T \tag{9.39}$$

$$I_{it}^{k}=I_{i,t-1}^{k}+Q_{it}^{k}Y_{it}^{k}-D_{it},\quad i=1,\ldots,N,\ t=1,\ldots,T \tag{9.40}$$

$$Q_{it}^{k}Y_{it}^{k}\le f_i\left(Q_{it}^{k},Y_{it}^{k},\overline{W}_{it}^{k},\widehat{\rho_{jt}^{k}},\widehat{\rho_{jt}^{k}T_{jt}^{k}}\right),\quad t=1,\ldots,T \tag{9.41}$$

$$Q_{it}^{k},Y_{it}^{k},I_{it}^{k},W_{it}^{k},R_{it}^{k},B_{it}^{k}\ge 0,\quad t=1,\ldots,T \tag{9.42}$$

where

$$\widehat{\rho_{jt}} = \sum_{j \neq i} \sum_{m \in \Upsilon_j} \tilde{\gamma}_{jt}^{k} Y_{jt}^{k} \left(s_j + p_j Q_{jt}^{k} \right) \tag{9.43}$$

denotes the utilization on the machine due to products other than i in the optimal solution to the restricted master problem at the current iteration. This pricing subproblem is a single-item dynamic lot-sizing problem, where the amount of capacity available to the product i is fixed by decisions corresponding to the other products. Dropping the constant $\propto_i^k$ we can write the objective function (9.38) as

$$\sum_{t=1}^{T} \left[h_{it} I_{it}^{k} + w_{it} W_{it}^{k} + b_{it} B_{it}^{k} + \alpha_{it}^{k} \left(s_i + p_i Q_{it}^{k} \right) Y_{it}^{k} \right] \tag{9.44}$$

and compare this to the objective function of the classical capacitated dynamic lot-sizing problem, which is given by:

$$\sum_{t=1}^{T} \left[h_{it} I_{it} + b_{it} B_{it} + S_{it} \Xi_{it} \right] \tag{9.45}$$

where Ξ_{it} is a binary variable equal to 1 if product i is produced in period t, and zero otherwise, while S_{it} denotes the fixed cost of a setup. Note that classical capacitated lot-sizing models all assume a single lot of each product in a given period, which would require the additional constraint Y_{it}^{k} ″ 1. Matching equivalent terms in (9.44) and (9.45) shows that the two objectives treat finished goods inventory and backlogs identically. Since the classical formulations do not consider congestion, and hence ignore WIP, let us assume that the cost of holding WIP is negligible. In this case, for (9.44) and (9.45) to give the same value, and hence the same solution, we must have

$$S_{it} = \alpha_{it}^{k} \left(s_i + p_i Q_{it}^{k} \right) \tag{9.46}$$

showing that even under the very restrictive assumptions imposed to achieve compatibility between the classical and MDCF-based lot-sizing models, the fixed cost of a setup must depend on the dual price of capacity at optimality—which is impossible to determine without obtaining an optimal solution for all products simultaneously.

Thus, while classical dynamic lot-sizing models can be justified in a purchasing environment, or in an environment with significant excess capacity, their use in a production environment is fraught with problems. Once the utilization of the machine reaches a certain point, it will become necessary to produce exactly one lot of each item in each period in which it is to be produced; however, the relative magnitudes of the fixed setup costs relative to inventory holding costs will determine the frequency of production. At lower utilization levels, however, (9.44) suggests that estimating the setup cost is far from trivial; at the very least, the setup cost for a product i will be time dependent, driven by the evolution of its demand over time as well as that of all other products competing with it for capacity.

9.4 Discussion

In this chapter, we have seen that the queueing perspective of Chap. 2 leads quite naturally to a series of models describing the impact of lot-sizing decisions on the performance of production units. The MDCFs developed in the previous chapter turn out to be a suitable mechanism to describe the behavior of such systems in mathematical programming models. The resulting optimization models are generally non-convex, requiring significant additional computational effort to guarantee a global optimal solution. However, there is considerable computational evidence that the non-convexity is of a somewhat benign nature; in many cases, the use of a convex nonlinear solver leads to confirmed global optimal solutions, suggesting the existence of considerable structure in the problem that remains an objective for future research. The use of steady-state queueing models to develop the MDCFs is clearly heuristic, and open to criticism; however, the significant improvements in system performance obtained in simulation experiments suggest that these models are worth developing further.

The contrast between these models and the traditional lot-sizing models that focus on the tradeoff between setup and holding costs is also informative. The (admittedly heuristic) column generation approach outlined in Sect. 9.3 highlights the complexity of estimating setup costs accurately. The results of Sect. 9.2, on the other hand, highlight a central implication of the traditional lot-sizing models such as the multilevel capacitated lot-sizing problem and its variants, which is that all production for a planning period must be produced in a single lot. Given that setup costs are charged on a per lot basis, this is natural, but the superior shop-floor performance obtained by the MDCF model suggests that especially at medium levels of utilization the use of smaller lot sizes can lead to considerable benefits. At high utilization levels there is no capacity to spare for additional setups, and hence the results of the traditional models and the MDCF model approach each other.

The work in this chapter is clearly exploratory in nature and merely scratches the surface of a broad and complex research agenda. The extension of this type of approach to multistage systems, such as those treated by Missbauer and Jutz (2018), or multilevel systems such as those arising in the context of MRP computations is a natural direction. It is unlikely that exact solutions to such formulation can be obtained for industrial scale problem instances, especially given the non-convex nature of many MDCF planning models, but an understanding of the structure of good solutions should serve as a pathway to computationally efficient approximations.

References

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

Benjaafar S (1996) On production batches, transfer batches and lead times. IIE Trans 28:357–362

Billington PJ, Mcclain JO, Thomas JL (1983) Mathematical programming approaches to capacity-constrained MRP systems: review, formulation and problem reduction. Manag Sci 29:1126–1141
Brahimi N, Dauzere-Peres S, Najid NM, Nordli A (2006) Single-item lot sizing problems. Eur J Oper Res 168:1–16
Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice-Hall, Englewood Cliffs
Cheng M, Mukherjee NJ, Sarin SC (2013) A review of lot streaming. Int J Prod Res 51(23/24):7023–7046
DeGraeve Z, Jans R (2007) A new Dantzig-Wolfe reformulation and branch and price algorithm for the capacitated lot sizing problem with setup times. Oper Res 55(5):909–920
Desaulniers G, Desrosiers J, Solomon MM (2005) Column generation. Springer, Berlin
Drexl A, Kimms A (1997) Lot sizing and scheduling—survey and extensions. Eur J Oper Res 99:221–235
Erenguc SS, Mercan M (1990) A multifamily dynamic lot-sizing model with coordinated replenishments. Nav Res Logist 37:539–558
Harris FW (1915) Operations and cost. In: Factory management series. A. W. Shaw Co, Chicago
Hopp WJ, Spearman ML (2008) Factory physics: foundations of manufacturing management. Irwin/McGraw-Hill, Boston
Jen Huei C, Huan Neng C (2005) A comprehensive review of lot streaming. Int J Prod Res 43(8):1515–1536
Jutz S (2017) Lot sizing in a two-stage production-inventory system—a flow time oriented perspective. Department of Information systems, production and logistics management. PhD, University of Innsbruck, Innsbruck, p 194
Kang Y, Albey E, Uzsoy R (2011) A column generation approach for multiple product dynamic lot-sizing with congestion. Raleigh, NC 27695-7906, Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University
Kang Y, Albey E, Hwang S, Uzsoy R (2014) The impact of lot-sizing in multiple product environments with congestion. J Manuf Syst 33(3):436–444
Kang Y, Albey E, Uzsoy R (2018) Rounding heuristics for multiple product dynamic lot-sizing in the presence of queueing behavior. Comput Oper Res 100:54–65
Karmarkar US (1987) Lot sizes, lead times and in-process inventories. Manag Sci 33(3):409–418
Karmarkar US (1989) Capacity loading and release planning with work-in-progress (WIP) and lead-times. J Manuf Oper Manag 2(1):105–123
Karmarkar US, Kekre S, Kekre S (1992) Multi-item batching heuristics for minimization of queues. Eur J Oper Res 58:99–111
Lasdon LS (1970) Optimization theory for large systems. Macmillan, New York
Lasdon LS, Terjung RC (1971) An efficient algorithm for multi-item scheduling. Oper Res 19(4):946–969
Medhi J (1991) Stochastic models in queuing theory. Academic, Cambridge
Missbauer H (2002) Lot sizing in workload control systems. Prod Plan Control 13:649–664
Missbauer H, Jutz S (2018) A flow time oriented lot sizing model for a serial two-stage production-inventory system: analytical approximation and simulation-based optimization. In: Twentieth international working seminar on production economics, Innsbruck, Austria
Pochet Y, Wolsey LA (2006) Production planning by mixed integer programming. Springer, New York
Quadt D, Kuhn H (2008) Capacitated lot sizing with extensions: a review. 4OR Q J Oper Res 6:61–83
Schneeweiss C (2003) Distributed decision making. Springer, Berlin
Söhner V, Schneeweiss C (1995) Hierarchically integrated lot size optimization. Eur J Oper Res 86(1):73–90
Trigeiro WW, Thomas LJ, McClain JO (1989) Capacitated lot sizing with setup times. Manag Sci 35:353–366

Vaughan TS (2006) Lot size effects on process lead time, lead time demand, and safety stock. Int J Prod Econ 100:1–9
Wijngaard J (1989) Timing and lot-sizing in production control. J Manuf Oper Manag 2:35–51
Zipkin PH (1986) Models for design and control of stochastic, multi-item batch production systems. Oper Res 34(1):91–104