Chapter 11
Conclusions and Future Directions

Problems arising in the PPC systems that support the complex global supply chains driving the modern economy were among the earliest to be addressed with the tools of operations research (Arrow et al. 1958; Hanssmann 1959; Holt et al. 1960; Buffa and Taubert 1972; Johnson and Montgomery 1974), leading to a broad, mature body of knowledge using a variety of mathematical formalisms including mathematical programming, queueing, simulation, and stochastic optimization. By the nature of this problem domain, this volume has ranged widely over a great deal of ground, and we hope that the reader has found the journey worthwhile. This chapter concludes the book with a brief review of the principal results and their implications for future work, both related to the clearing functions that are the central concern of this volume and for the broader field of production planning models.

11.1 The Gordian Knot: Output, Cycle Time, and Workload

The problem at the heart of this volume is the intimate interconnection between the output, cycle time, and workload of a production unit, with, of course, the individual production resource as a special case. Whether the problem faced is that of coordinating a number of production units across a supply chain, or planning the releases of work into an individual production unit to meet demand in the best possible manner, PPC systems simply cannot operate effectively without some cognizance of the impact of their decisions on cycle times. As discussed in Chap. 2, queueing models, simulation experiments, and industrial observation all indicate that the cycle time of an order through a production unit is a random variable whose distribution depends, among potentially many other things, on the utilization of the resource, i.e., the workload available for it to process that is determined by the work release decisions made by any PPC system discussed in Chap. 1. Hence cycle times should be considered as endogenous to the planning process, rather than as an exogenous parameter, which is manifestly not the case in most of the production planning literature that

H. Missbauer, R. Uzsoy, *Production Planning with Capacitated Resources and Congestion*, https://doi.org/10.1007/978-1-0716-0354-3_11

goes by that name. Most of this literature can, with more or less of a stretch, be positioned within one of the two principal frameworks for PPC systems that have emerged as a (admittedly evolving) consensus between industrial practice and academic research: Manufacturing Planning and Control (Jacobs et al. 2011) and Advanced Planning and Scheduling (APS) (Stadtler et al. 2015). The discussion of these frameworks in Chap. 3 highlights the importance of cycle times to the effective operation of both. This circularity—that planning systems need to be cognizant of cycle times, but cycle times are a consequence of the work release decisions made by the planning systems themselves—has, in our opinion, constituted a significant barrier to progress. The contents of this volume can thus be viewed as a series of attempts to address this difficulty or mitigate its negative impacts.

Chapter 4 discusses the workload control (WLC) paradigm, which constitutes a first-order response to the relation between workload, output, and cycle time. Despite their wide variety, all WLC approaches seek to identify a workload level for the production unit that will yield an acceptable compromise between the goals of maintaining low WIP and cycle times on the one hand, and sufficient output to meet demand on the other. Most such systems are rule-based, designed to operate in an environment where the demand distribution faced by the production unit remains approximately constant; they do not easily adapt to changing operating conditions, which would require recalculation of their various parameters as the environment changes. It is probably fair to say that there is as yet no unified theory governing the relations between the environmental conditions faced by such WLC systems and the values of the various parameters they require. Only a few of these approaches use an explicit model of material flow through the production unit to inform their work release decisions. The optimization models discussed in the subsequent chapters can be viewed as natural extensions of these model-based WLC approaches.

Most existing approaches to production planning, from the material requirements planning (MRP) procedure widely used in industry (Orlicky 1975; Baker 1993; Jacobs et al. 2011) to the mathematical programming models that form the central engine of many advanced planning and scheduling (APS) systems (Voss and Woodruff 2006; Hackman 2008), approach this issue using planned lead times that are treated as exogenous, workload-independent parameters. As long as lot-sizing or capacity expansion decisions are not considered, avoiding the need for integer variables, these models can generally be formulated as linear programs that can be solved with existing commercial solvers, even for very large problem instances. One of us (RU) had the opportunity a decade ago to observe the implementation of a new planning system at a major high-technology manufacturer. The complete workflow for generating a plan for a significant portion of the supply chain, involving multiple plants, multiple production lines within plants and distribution facilities, took approximately 24 h at that time, of which only 45 min was required for the solution of the optimization model. The remaining time was taken up by acquiring, formatting, and cleaning input data from the firm's ERP system and then transferring the output of the planning model back to the ERP system for execution. Chapter 5 summarizes the state of the art in these models when the planned lead time remains constant over time.

There is considerable evidence, including our own presented in Chap. 10, that despite its evident inconsistency with queueing theory the use of fixed planned lead times frequently does not lead to unacceptably bad performance. One reason for this may be that many facilities are operated within a relatively narrow range of operating conditions as defined by product mix, available resources, and demand, allowing planned lead times that provide good shop-floor performance to be arrived at over time. There is also often considerable opportunity for shop-floor decisions to mitigate the negative effects of suboptimal work release decisions by scheduling overtime, exploiting alternative resources, expediting and other such measures.

An additional advantage of planning models based on exogenous planned lead times is their intuitive nature. The idea of a delay between the release of work and its emergence as finished product is an easy one to grasp, making acceptance of the resulting planning models by their ultimate users, the managers responsible for the performance of the production units making up the supply chain, much easier than for a complex, nonlinear mathematical model. This does not mean, however, that the decisions obtained from a complex optimization model are always intuitive; anyone who has tried to explain to a manager why the optimization model chose to produce a specific amount of a specific item at a specific time on a specific resource, instead of using one of the many available alternatives (usually including the manager's favorite), will recognize the difficulty in parsing the output of a large mathematical program into a narrative explanation. The work of Greenberg (1996) on a rule-based system for explaining the results of linear programming models suggests an interesting direction for future research customizing this generic approach to specific production planning formulations.

The planning models in Chap. 5 can be viewed as optimizing work releases for a given set of planned lead times. The endogeneity of cycle times to work release decisions discussed in Chap. 2 suggests a model that can jointly optimize releases and cycle times simultaneously. Thus, if we could find the "correct" planned lead times for each planning period, the models of Chap. 5 would provide the optimal releases directly. Chapter 6 explores the difficulties that arise in identifying a consistent set of planned lead times across the planning horizon, and then focuses on approaches that decompose the planning problem into two subproblems. The first of these takes estimates of planned lead times as input and computes optimal releases based on these lead times. The second model takes a set of releases as input, and returns estimates of the resulting cycle times from which revised planned lead times can be computed. The release planning model is usually a linear program similar to those described in Chap. 5, while the lead time estimation model is usually a more or less detailed simulation model of the production unit of interest, although queueing and statistical models can also be used. A variety of such models have been proposed since the initial work of Hung and Leachman (1996), none of which have yielded conclusively positive results. Their computational burden tends to be high due to the need for multiple replications of a (often large) simulation model at each iteration. Their convergence behavior is not well understood; there appears to be no theoretical guarantee of their convergence, and experimental observations include cycling between solutions, failure to converge in any recognizable way, and dependence on the starting solution. The prime advantage of these approaches is that they combine two techniques, mathematical programming and simulation, that are each familiar to practitioners and have access to excellent commercial software. However, this approach does not build an explicit model linking output, workload, and cycle time; this model is implicit in the dynamics of the simulation or queueing model used to estimate the planned lead times given a set of releases.

Chapter 7 introduces univariate clearing functions that formulate a mathematical relation linking the expected output of a production resource in a planning period to some measure of its workload in the planning period. The basic concept was introduced, apparently independently, by several researchers in the late 1980s (Graves 1986; Srinivasan et al. 1988; Karmarkar 1989). Univariate clearing functions that are concave in their measure of workload (whatever that may be) yield convex optimization models. Since a concave function can be approximated to any degree of accuracy by a set of linear functions, it is easy to approximate these as linear programs, although the growing computational power of convex nonlinear solvers renders this less important than it once was. However, serious difficulties arise when multiple products competing for capacity on the same resource are considered, and straightforward extension of the single-product models results in clearly anomalous behavior. These difficulties are closely related to those ably explored by Carey and his coauthors in the domain of traffic modeling (Carey 1987, 1990; Carey and Subrahmanian 2000; Carey and Bowers 2012) and discussed in Chap. 6 in the context of time-varying planned lead times. After illustrating the behavior of the clearing function as a representation of a production unit, this chapter presents the allocated clearing function model of Asmundsson et al. (2006, 2009), which provides an effective although approximate solution to these difficulties and remains the state of the art at this time of writing. The chapter also illustrates one of the primary theoretical advantages of the clearing function approach over the models of Chap. 5, its ability to provide richer dual information on the marginal price of capacity at the different resources in the production unit.

The development of the allocated clearing function model exposes the limitations of the use of univariate clearing functions. The univariate clearing function estimates the aggregate output of the production unit across all products as a function of the aggregate workload of all products and then, as its name implies, allocates this aggregate output optimally among the different products. Chapter 8 departs from the observation that the allocated clearing function approach fails quite badly when the aggregate output depends heavily on the mix of products, not just on the aggregate workload. This is clearly the case when there are significant setup times between different products on the production resources; lot-sizing and sequencing decisions now have major impact on output. This chapter examines efforts to formulate multivariate clearing functions, raising the question of what additional state variables should be included. A variety of such state variables have been tried, including decomposing the workload of a product into the WIP available at the start of the period and releases during the period; inclusion of state variables related to previous periods; and using the output of each product as a state variable describing the output of all others. Many of these efforts result in non-convex optimization models, although computational evidence suggests that convex solvers can often obtain global optimal solutions in many cases, suggesting the presence of considerable structure that remains to be uncovered. Computational results, however, indicate that considerable improvement over univariate clearing functions can be obtained, at the cost of additional computational burden. It is probably safe to classify much of the work in this chapter as exploratory, leaving considerable room for future research.

Chapter 9 briefly explores the relation between the clearing function concept and lot-sizing decisions in the context of a single production resource. The seminal work of Karmarkar and his coworkers (Karmarkar 1987; Karmarkar et al. 1992) used queueing models to illustrate the relation between lot-sizing decisions and cycle times, which can then be used to derive multivariate clearing functions in which the output of a product depends on the lot sizes and output of all products in the system. The chapter then develops a non-convex optimization model using multivariate clearing functions for a single-machine dynamic lot-sizing problem and shows that this can yield significant performance improvements over prior approaches that do not consider queueing behavior. The chapter closes with an admittedly heuristic discussion using this model to illustrate the difficulty of accurately estimating the setup costs that are a crucial parameter of most lot-sizing models in the literature, which focus on the tradeoff between setup and cycle stock holding costs.

Having presented the clearing function concept in various forms in Chaps. 7 through 9, Chap. 10 examines several applications of the concept. A series of computational experiments using the allocated clearing function model for release planning for semiconductor wafer fabrication yield admittedly mixed results. While the clearing function model outperforms fixed lead time models with integer lead times, the use of fractional lead times largely eliminates the advantage of the clearing functions except under time-varying demand. Other applications include the use of clearing function models in a rolling horizon context, where they largely retain their advantage over fixed lead time models, the integrated planning of production and improvement activities, and dynamic pricing in an environment where demand is sensitive to both lead time and price. By and large, the results of the clearing function approaches are promising, especially when the richer dual information they yield can be used to gain insight into system behavior.

11.2 Weaknesses and Limitations of the Clearing Function Approach

Having laid out in the preceding chapters the basic motivation for the clearing function approach and the state of our knowledge to date, we would be remiss if we implied that we have a watertight case; we most certainly do not. The perceptive reader will have raised a number of criticisms themselves by this point in the volume, and there are many such both stated and implied in the previous pages. In this section, we will discuss several of the most important of these difficulties, some of which are the subject of ongoing research while others await the attention of the research community.

11.2.1 Why Clearing Functions?

The first question that needs to be addressed is simply that of why the clearing function construct should be used at all. Chapter 7 pointed out that a clearing function is a metamodel describing certain aspects of the behavior of a queue. If this is indeed the case, there are a wide range of alternative approaches to choose from, such as other forms of metamodeling (Li et al. 2016), system dynamics (Sterman 2000), transient queueing models (Askin and Hanumantha 2018) and, of course, simulation (Law and Kelton 2004). Given that the reason for using clearing functions is not at all obvious, and at least one reviewer of our work has stated categorically that "… the clearing function idea is outdated," some discussion of this issue appears to be necessary.

The primary reason for using a clearing function to represent a production unit is the ability to embed it in a tractable optimization model to plan releases for the next several periods. For this purpose, what is required is a sufficiently accurate representation of the relation between workload and output; six decimal places of precision are not required in a planning model whose purpose is simply to ensure that the workload in the production unit is at the correct level to meet the desired output without unnecessarily increasing WIP and cycle time.

The other types of model described above do not lend themselves easily to the formulation of tractable mathematical programming models. It is certainly true that queueing or simulation models can be embedded in an optimization framework, using algorithms similar to those used for simulation optimization (Fu 2015). A number of models of this latter type have been presented in the literature, notably the metamodel-based simulation optimization algorithm of Li et al. (2016) and the simulation optimization approaches of Kacar and Uzsoy (2015). These models benefit from the superior ability of simulation models to incorporate detailed system dynamics that are difficult to capture in a clearing function. However, the development of the metamodel requires extensive simulation experiments to collect data and fit the metamodel, while simulation optimization is very time-consuming. The use of the clearing function construct is aimed at enabling the use of a mathematical programming model to optimize releases as well as providing the information from the dual solution that may help management better understand the behavior of their system. It is very unlikely that a clearing function can provide a highly precise prediction of output in each period, but that is not its purpose; it seeks to provide sufficiently accurate descriptions of system behavior to ensure that the planning model using it maintains the system workload in a state that will sustain the desired output level.

The advantage of a mathematical programming model, in turn, is that it allows rapid solution of a complex optimization problem, especially when it can be formulated as a linear program. The estimation of the clearing function used in the model will require considerable computational effort, but this work can be performed offline and is not part of the run generating the planning solution. In contrast, any model utilizing a detailed simulation of the production unit of interest, whether a full-blown simulation optimization approach or one of the iterative multi-model approaches discussed in Chap. 6, requires multiple replications of simulation runs for the actual solution of the release planning model, requiring considerably more time.

11.2.2 Choice of Functional Form for the Clearing Function

While the idea of a concave non-decreasing functional relation between the workload and the expected output of a production resource is quite intuitive, it should be evident to the reader by this point that the state of our knowledge as to what functional forms to use and how to estimate their parameters from either industrial or simulation data is as yet highly unsatisfactory. The derivation of clearing functions from steady-state queueing models is inherently dangerous in a discrete-time planning model unless planning periods are long enough for the underlying queues to at least approximately reach steady state, which is frequently not the case in practice. The early work of Asmundsson et al. (2009) revealed that using conventional least-squares regression to fit one of the empirical functional forms discussed in Chap. 7 results in systematic overestimation of the expected output, due to the relation captured by Jensen's inequality (8.8) discussed in Sect. 8.2. Gopalswamy and Uzsoy (2019) identify a number of additional issues arising in the fitting of clearing functions to simulation data, and the design of appropriate simulation experiments to obtain such data.

Even if the issues associated with estimating clearing functions of a tractable computational form were addressed satisfactorily, the currently common approach of fitting a single clearing function that is expected to represent the behavior of the production resource in all planning periods is clearly a significant approximation, as discussed in Sect. 8.2 and illustrated in Fig. 8.2. The simulation optimization approach of Kacar and Uzsoy (2015) found that fitting a clearing function to each planning period gave superior results to using a single clearing function for all periods. However, this simulation optimization approach is computationally demanding, especially when used in a rolling horizon environment.

Yet another difficulty with the use of clearing functions arises in multistage environments. Expression (2.1) shows that the expected cycle time at a given resource depends on the mean and variance of both interarrival and service times; the variability of the interarrival times in turn depends on decisions made at upstream resources. Thus, at least in theory, the shape of the clearing function at a given resource is affected by the production decisions at upstream resources. The clearing function models we have discussed in this volume do not take this type of interrelation between production resources, or production units, into account. Instead they assume that the mean and variance of interarrival and service times at each resource are independent of other resources, an assumption referred to in queueing theory as decomposition. Clearly some error is introduced into the models by this approach. One would expect this to become especially serious in multistage systems with setup times at each resource, where lot-sizing decisions at each stage in each period may affect the shape of clearing functions at downstream resources.

The situation for MDCFs is still more complicated. The fact of the matter is that at present we have no firm theoretical foundation for deciding which state variables to include in an MDCF; it is notable, and lamentable, that most of the MDCFs proposed to date draw their functional form from steady-state queueing analyses, often of very simple models. For example, the MDCFs of Albey et al. (2014, 2017) follow the functional form suggested by Karmarkar (1989) which is motivated primarily by steady-state analysis of the *M/M/1* queue. The experimental work of Gopalswamy and Uzsoy (2019) suggests that the empirical functional forms used extensively in the past cannot provide good fits across the entire operating range of workloads a production resource will encounter.

Our current state of knowledge suggests that the best approach to fitting clearing functions available at present is the use of concave piecewise linear regression, which can be formulated as a mixed integer program (Toriello and Vielma 2012; Gopalswamy et al. 2019) although the solution of large models with many data points remains computationally challenging. The piecewise linear approach allows great modeling flexibility and yields a clearing function that when implemented in the allocated clearing function model of Chap. 7 results in a linear program. However, the establishment of a strong theoretical and methodological foundation for the fitting of clearing functions, encompassing both the choice of state variables and of a suitable functional form, remains important directions for future research. The promising performance of clearing function based production planning models presented in this volume suggests that this effort may well be worthwhile.

11.3 Some Directions for Future Research

The limitations of the clearing function approach discussed in the previous section suggest a broad range of interesting research questions for the future, many of which lie at the intersection of what have traditionally been viewed as quite distinct research streams. The basic idea of a clearing function lies at the intersection of queueing and mathematical programming models of production systems, research areas that have developed largely independently until today. In this section, we discuss several longer-term research efforts that can build on the clearing function concept, but which address much broader issues spanning several research streams and mathematical modeling tools.

It is clear from the discussion in Chaps. 5 and 6 of this volume that, from a technical perspective, order release models and mechanisms that assume fixed, exogenous lead times are quite mature, although their integration into the overall PPC system can raise numerous questions. Load-dependent lead times, on the other hand, are much more difficult to handle technically (and also organizationally, although this is not our primary focus), and this research stream is far from mature. This provides great opportunities for researchers to advance the frontier of our knowledge in this domain that is, as described in Chap. 1, an essential element of the PPC architecture in most discrete manufacturing companies. We now briefly describe some of the most important research questions, starting with technical issues and proceeding to more conceptual topics.

11.3.1 Parameter Setting for Order Release Models

Order release models with exogenous lead times require lead time parameters that are often taken to be constant over time as in Chap. 5, but can also vary over time as discussed in Chap. 6. Clearing function models must specify the functional form and the shape parameters of the clearing functions. These parameters anticipate the behavior of the production units, but since both the realized cycle times and the realized output are random variables subject to often unknown and changing probability distributions, the parameters cannot simply be set to the "correct" values. This is especially evident for clearing functions where the conditional distribution of the output for a given planned load depends on various factors including the order release pattern itself, due to the planning circularity described in Chap. 2. Therefore, the choice of parameter values encompasses both an anticipation aspect (how accurately will the clearing function anticipate the realized output from the production unit or resource?) and an implicit decision as to the tradeoffs between WIP and FGI inventory levels and due date performance (with the importance of the latter depending on whether safety stocks are maintained, as discussed below). The performance of order release models can be quite sensitive to the parameter values as indicated, e.g., by the performance differences between fixed lead time models with integer and fractional lead times discussed in Chaps. 5 and 10. Very similar questions can be raised in terms of estimating suitable fixed lead times: assuming the distribution of the cycle times in each planning period was known, what is the optimal value of the planned lead times?

Considering the anticipation aspect of the parameter setting problem, one would assume that best performance can be achieved by setting, e.g., clearing function parameters to the values obtained from observation, such as running a least-squares regression over observed load–output data. However, the parameter setting that yields the best system performance can be substantially different (Kacar and Uzsoy 2015) and the mechanisms behind these deviations are not fully understood. We must also keep in mind that the vast majority of research on parameterization issues is performed on simulated data. Empirical data exhibit substantial noise which makes functional relationships between load and output difficult to identify (Fine and Graves 1989; Häussler and Missbauer 2014), and further research is needed to examine the validity of insights obtained from simulations to real-life situations.

11.3.2 A Deeper Understanding of Clearing Functions: Properties, Theoretical Basis, and Integration with Order Release Models

In Chap. 7, clearing functions were motivated by queueing models that suggest a concave, saturating functional relationship between WIP and output, caused primarily by the variability of the arrival and departure process. However, the clearing function concept was introduced by Graves (1986) assuming that output in a period is proportional to the load in that period. The smoothing parameter that is implied by this model is assumed to be "set … so that the resulting time series for production is consistent with the work center capability" which can be obtained by assuming that "As a queue builds at a work center, a manager will direct more resources to the work center to reduce the queue to normal levels" (p. 524). Hence this proportional clearing function models the effect of a production control rule, which is quite distinct from the variability argument invoked to justify the nonlinear, saturating shape. Linear and saturating clearing functions differ not just with respect to their shape, but also with respect to the underlying phenomenon they seek to represent. It is important to keep both modeling aspects of clearing functions in mind—modeling variability versus modeling production control rules. The latter aspect opens up the possibility of modeling behavioral aspects such as load-dependent processing times, possible capacity loss due to congestion (e.g., because material must be shuffled by the production workers), etc. The modeling of these often largely informal factors is still at its beginning—an important research question within behavioral operations management—and can substantially influence the behavior of clearing function models.

Applying clearing functions in a transient regime leads to additional complications. While it is easy to prove that decomposing the workload in a period t (the explanatory variable of most one-dimensional clearing functions) into its components and formulating a multi-dimensional clearing function that takes the history of the process into account can improve output estimation, incorporating this function into order release models can lead to oscillating order releases as discussed in Chap. 8. Naïvely we would assume that more accurate anticipation of the output should improve the performance of the optimization model, but apparently things are not that simple. This indicates that a comprehensive, consistent theory of order release models incorporating functions that estimate the conditional future output is not yet available. This aspect also raises the question of which characteristics of queueing systems are most critical to the model behavior and thus should be modeled most accurately. These might include the steady-state behavior, the WIP and output evolution in the transient phase, or the propagation of variability through the workcenters, including the transient phase.

Production orders that are released are eventually finished, unless they are canceled deliberately; only their finish time is uncertain. Clearing function models express this timing uncertainty as an uncertainty in output quantities across the periods. This maintains the basic structure of production planning models established in the pioneering works described in Sect. 4.6, and is a significant modeling decision since this basic structure was not originally designed for handling lead times. Integrating lead time variables into this modeling framework, that is, expressing the timing uncertainty directly, either leads to intractable model structures or discards the tight relationship between WIP and cycle time demonstrated in Chap. 2. The difficulty with cycle time oriented release models and iterative approaches is an immediate consequence. Research on alternative modeling approaches such as robust optimization is an obvious possibility.

11.3.3 Integration of Order Release into the Overall Supply Chain

Both model-based and rule-based order release mechanisms mainly deal with order releases to single production units. This is adequate for MTO companies where customer orders translate directly to production orders in the order pool and are processed mainly by a single production unit, like the CD/DVD manufacturer in Sect. 1.2.2. However, if the orders must be processed sequentially by multiple production units as in semiconductor manufacturing (Sect. 1.2.1), the order releases to the production units must be coordinated according to the BOM structure. Fixed lead times greatly simplify this material coordination task (de Kok and Fransoo 2003) and can, in principle, be combined with clearing functions (Jansen et al. 2013). Extending load-dependent lead time models to incorporate material coordination along the multistage production-inventory system is much more difficult and remains a topic for future research. A central question is whether a sequential approach that derives the demand for the end items of a production unit from the planned releases to the downstream production units, or an integrated model that encompasses all production units simultaneously, extending the fixed lead time supply chain model in de Kok and Fransoo (2003), is preferable. This problem can be viewed, at an extreme, as that of incorporating the queueing behavior of resources described in Chap. 2 into the MRP logic of the MPC framework described in Chap. 3; should we modify the release schedule obtained by the MRP logic after the fact to accommodate the effects of limited capacity and queueing, or should this be done within the MRP run itself in some way? Analogously, if master planning as implemented in APS systems is applied the queueing perspective must be integrated into the capacity and lead time modeling used at this level.

Order release models integrate the tasks of production smoothing and cycle time anticipation and control. Production smoothing is also performed at the master planning/MPS level to ensure that the number of end items requested from a production unit in each period (the D_{jt} parameters in the release models) is consistent with the capabilities of the production unit. Hence the smoothing capabilities required at the order release stage depend on the smoothing logic at the master planning/MPS level, and perhaps even at the lot-sizing level. Seamless integration of these levels requires consistency between the decision models at each level, in particular in how they anticipate the dynamic behavior of the production units. The anticipation models applied at the master planning/MPS level should be aggregate versions of the respective models applied for order release. Since master planning models and resource profiles for master production scheduling are generally not WIP based, this is not trivial. The research task behind this is the aggregation of transient queueing networks. Finding suitable approximations that are applicable in practice is still largely unsolved; Zäpfel and Missbauer (1993) give a first attempt to handle this aggregation problem. To what extent aggregation methods designed for the steady state, such as the effective processing time approach of Kock et al. (2011), Hopp and Spearman (2008), and Veeger et al. (2011), can be extended to the transient states implied by load-dependent lead times remains to be clarified.

Integrating order release models into the material coordination task also requires the determination of inventory levels of stock keeping units between the production units, including safety stocks. Order release models that explicitly determine WIP levels within the production units provide, at least in principle, the possibility of considering the interaction between WIP and safety stock that is evident from inventory theory—the stockout probability for a given demand distribution depends on both the safety stock and the WIP level, with WIP providing some functionality of safety stock (Graves 1988). Simultaneous optimization of safety stocks and lead times or planned WIP, respectively, using clearing function models and an aggregate representation of the production units has been demonstrated in Albey et al. (2015), Albey and Uzsoy (2016), Aouam and Uzsoy (2012, 2015), and Orcun et al. (2009). Extending this work to multistage systems and/or more detailed representations of the production units is an obvious research topic.

Optimizing inventory levels refers to handling uncertainty of demand and other planning parameters. Since demand uncertainty is usually higher for more remote planning periods, which implies that for a certain period it becomes smaller as the time of planning proceeds, simply optimizing the (safety) stock levels for given demand distributions can lead to exaggerated planned stock levels for the more distant planning periods that most likely are corrected in the course of rolling horizon planning; the order release and capacity plans are biased systematically. Stochastic demand and specific planning rules for responding to demand render future production quantities random variables (de Kok and Fransoo 2003), as considered in the original work of Holt et al. (1960). Aouam and Uzsoy (2012) find that in their simple setting linear decision rules for updating the planned production quantities perform well, which is an encouraging result and in line with adjustable robust optimization (Gorissen et al. 2015). How to extend this to realistic settings and how to assign/split up the task of setting planned inventory levels to/between the order release and the master planning/MPS level is a challenging research topic since it encompasses the hierarchical design of the entire PPC system. Again the need to consider the impact of planned WIP at the master planning level is evident.

11.3.4 Advanced Techniques for Flow Time and Output Modeling

Determining lead times and order release quantities simultaneously requires an anticipation model that predicts the future values of these performance measures for a given release schedule. Univariate clearing functions consider only some measure of WIP as explanatory variable, while multi-dimensional clearing functions (MCDFs) allow more accurate representation of the causal mechanisms that lead to certain values for cycle times and output. The conceptual problems of MDCFs, especially when transient effects are modeled, are described in Chap. 8. Very accurate cycle time and output prediction can be obtained by discrete-event simulation, but this leads to the difficulties described in Sect. 6.6.

This dilemma motivates the use of metamodeling for cycle time and output prediction and metamodel-based rather than simulation-based optimization (Barton and Meckesheimer 2006). When appropriately trained or parameterized, the metamodel, which is usually a deterministic function, yields estimates of the performance measures of the production unit very close to the simulation output as a function of the input variables, which in our case are the order releases over time. The impact of the relevant parameters that describe the properties of the material flow, such as machine failure characteristics, lots sizes, and operation times of the production lots, are either coded in the metamodel or are declared as arguments of the metamodel depending on its specification. Metamodels can be represented by generic functions such as polynomials, functions that are based on certain theoretical requirements on their shape (e.g., the MDCF in Häussler and Missbauer 2014), or by artificial neural networks. Applying metamodels to anticipate output and cycle times is an extension of MDCFs. Based on Yang and Liu (2012) who propose a metamodel for the transient analysis of queueing systems, Li et al. (2016) develop a metamodel that receives the release quantities in the planning periods as input and yields the first two moments of WIP and output in the planning periods. Given this metamodel, the releases are approximately optimized using a multi-objective genetic algorithm. The metamodel considers both the departures and the queue length over the relevant past periods for the output prediction, making reasonable assumptions about the underlying time series model, and is fitted using extensive simulation data. The model performs very well compared to a simulation-based optimization with excessively long computer run times. Since there is no sharp boundary between MDCFs and metamodels, there might be a number of ways to formulate and refine metamodels for anticipating output and cycle times that can be explored in future research.

In the metamodeling approach, the release decision is decomposed into the two phases of pre-computing the metamodel and optimizing the decision variables, and the metamodel is fitted which usually requires large amounts of data that normally can only be obtained by simulation. Both of these aspects raise difficulties: the decomposition into metamodeling and optimization might result in suboptimal solutions, and the effects of informal shop-floor control rules might be difficult to capture in a simulation model. This motivates the use of machine learning techniques that strictly learn from observed data, either to learn the response of the production unit to control inputs (e.g., releases) or to learn near-optimal control inputs for a given state of the system directly (for this distinction, see Bertsimas et al. 2019). While the application of machine learning at the scheduling level has been explored extensively (Aytug et al. 2005), very few papers apply machine learning at the order release level. Lee et al. (1997) use machine learning to select the release sequencing rule in a CONWIP system. Paternina-Arboleda and Das (2001) use reinforcement learning to optimize the operation of an extended CONWIP system which also constrains the buffers at the workcenters and allows emergency authorizations of releases, similar to the force release option in LUMS (see Chap. 4). Häussler and Schneckenreither (2019) use an artificial neural network to predict the cycle time of a new order entering the production unit and, based on this estimation, determine the release times of the production orders, thus decomposing the problem of jointly determining the release times of the orders to single-order release problems that are combined by an algorithm developed in the paper. Clearly these approaches are first attempts, and further research in this area seems fruitful.

11.4 Conclusions

The domain of production planning is viewed by many as a mature area where all interesting problems have already been solved. We hope that the results presented in this volume have raised more questions in the mind of the reader than they have answered; this has been the effect of this work on the authors, in any event. There remain many challenging problems that, if even approximately solved, have the potential to yield significant economic benefit to many sectors of the economy. The convergence of vast computing power, data collection and storage technologies and extremely efficient optimization solvers, as well as developments in data analytics, stochastic optimization and machine learning, open new possibilities for advances in this area which has, after all, been central to the development of operations research, operations management, production economics, and industrial engineering since the inception of those disciplines. It is, we believe, a good time to be working in production planning and will only get better.

References

Albey E, Uzsoy R (2016) A chance constraint based multi-item production planning model using simulation optimization. In: Winter simulation conference, Arlington, VA

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

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

Arrow KJ, Karlin S, Scarf H (1958) Studies in the mathematical theory of inventory and production. Stanford University Press, Stanford

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

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. Nav Res Logist 56(2):142–157

Aytug H, Lawley MA, McKay KN, Mohan S, Uzsoy R (2005) Executing production schedules in the face of uncertainty: a review and some future directions. Eur J Oper Res 161:86–110

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

Barton RR, Meckesheimer M (2006) Chapter 18: Metamodel-based simulation optimization. In: Henderson SG, Nelson BL (eds) Handbooks in operations research and management science, vol 13. Elsevier, Amsterdam, pp 535–574

Bertsimas D, Dunn J, Mundru N (2019) Optimal prescriptive trees. INFORMS Journal on Optimization 1(2):164–183

Buffa ES, Taubert WH (1972) Production-inventory systems; planning and control. R. D. Irwin, Homewood

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

Carey M (1990) Extending and solving a multiperiod congested network flow model. Comput Oper Res 17(5):495–507

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

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

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

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

Fu MC (ed) (2015) Handbook of simulation optimization. Springer, New York

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

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

Gorissen BL, Yanikoglu I, den Hertog D (2015) A practical guide to robust optimization. Omega 53:124–137

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

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

Greenberg HJ (1996) The analyze rulebase for supporting LP analysis. Ann Oper Res 65:91–126

Hackman ST (2008) Production economics: integrating the microeconomic and engineering perspectives. Springer, Berlin

Hanssmann F (1959) Optimal inventory location and control in production and distribution networks. Oper Res 7(4):483–498

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

Häussler S, Schneckenreither M (2019) Adaptive order release planning with dynamic lead times: a machine learning approach. University of Innsbruck, Innsbruck

Holt CC, Modigliani F, Muth JF, Simon HA (1960) Planning production, inventories and work force. Prentice Hall, Englewood Cliffs

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

Hung YF, Leachman RC (1996) A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations. IEEE Trans Semicond Manuf 9(2):257–269

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

Jansen MM, de Kok TG, Fransoo JC (2013) Lead time anticipation in supply chain operations planning. OR Spectr 35(1):251–290

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

Kacar NB, Uzsoy R (2015) Estimating clearing functions for production resources using simulation optimization. IEEE Trans Autom Sci Eng 12(2):539–552

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

Kock AAA, Veeger CPL, Etman LFP, Lemmen B, Rooda JE (2011) Lumped parameter modelling of the litho cell. Prod Plan Control 22(1):41–49

Law AM, Kelton WD (2004) Simulation modeling and analysis. McGraw-Hill, New York

Lee CY, Piramuthu S, Tsai YK (1997) Job shop scheduling with a genetic algorithm and machine learning. Int J Prod Res 35(4):1171–1191

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

Orcun S, Uzsoy R, Kempf KG (2009) An integrated production planning model with load-dependent lead-times and safety stocks. Comput Chem Eng 33(12):2159–2163

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

Paternina-Arboleda CD, Das TK (2001) Intelligent dynamic control policies for serial production lines. IIE Trans 33(1):65–77

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

Stadtler H, Kilger C, Meyr H (2015) Supply chain management and advanced planning. Concepts, models, software, and case studies. Springer, Berlin

Sterman JD (2000) Business dynamics: systems thinking and modeling for a complex world. McGraw-Hill, New York

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

Veeger CPL, Etman LFP, Lefeber E, Adan IJBF, van Herk J, Rooda JE (2011) Predicting cycle time distributions for integrated processing workstations: an aggregate modeling approach. IEEE Trans Semicond Manuf 24(2):223–236

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

Yang F, Liu J (2012) Simulation-based transfer function modeling for transient analysis of general queueing systems. Eur J Oper Res 223(1):150–166

Zäpfel G, Missbauer H (1993) Production planning and control (PPC) systems including load-oriented order release—problems and research perspectives. Int J Prod Econ 30/31:107–122