This article was downloaded by: [216.73.216.247] On: 14 November 2026, At: 09:25
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA



INFORMS TutORials in Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Continuous Dynamic Models, Clearing Functions, and Discrete-Event Simulation in Aggregate Production Planning

Dieter ArmbrusterReha Uzsoy

To cite this entry: Dieter ArmbrusterReha Uzsoy. Continuous Dynamic Models, Clearing Functions, and Discrete-Event Simulation in Aggregate Production Planning. *In* INFORMS TutORials in Operations Research. Published online: 14 Oct 2014; 103-126.
https://doi.org/10.1287/educ.1120.0102

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2012, INFORMS

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

TUTORIALS IN
OPERATIONS RESEARCH
INFORMS 2012

© 2012 INFORMS | ISBN 978-0-9843378-3-5
http://dx.doi.org/10.1287/educ.1120.0102

Continuous Dynamic Models, Clearing Functions, and Discrete-Event Simulation in Aggregate Production Planning

Dieter Armbruster
School of Mathematical and Statistical Sciences, Arizona State University, Tempe, Arizona 85287, armbruster@asu.edu

Reha Uzsoy
Edward P. Fitts Department of Industrial and Systems Engineering,
North Carolina State University, Raleigh, North Carolina 27695, ruzsoy@ncsu.edu

Abstract We consider the problem of representing the capabilities of production systems to convert inputs into outputs over time in the face of deterministic demand. The fact that planning models generally operate in discrete-time while the factory being modeled operates in continuous time suggests the use of multimodel methods combining an optimization model that plans the releases into the plant over time with a simulation model that assesses the impacts of these releases on performance. We discuss several such schemes that have been implemented using discrete-event simulation, present recent computational results, and discuss their strengths and weaknesses. We then consider two alternative approaches that have emerged in the recent literature: the use of discrete-time linear programming models based on nonlinear clearing functions, and methods using systems of coupled partial differential equations based on transport phenomena. We identify the relationships between queueing models, clearing functions, and transport model-based methods, and we discuss future research directions.

Keywords production planning; clearing functions; linear programming; transport equations; partial differential equations

1. Introduction

The problems of planning and controlling production in the manufacturing industries have been a key application domain for industrial engineering and operations research since the emergence of these fields as recognized disciplines. The extensive body of literature in these areas dates back to at least the beginning of the 20th century, including the work of Harris [34] and Arrow et al. [11] on inventory models, and that of Modigliani and Hohn [60] and Holt et al. [35] on discrete-time production planning. The problem of production planning we shall address in this tutorial addresses the by now classical formulation of the latter group of authors. In its simplest form, a firm operates in a market where it faces external demand, which it tries to meet with a limited set of production resources with limited ability to generate output in a given time period. The limitation on the amount of output that can be produced in a given time interval by a production resource is colloquially referred to as its capacity. The precise definition and determination of capacity turns out to be a challenging problem, as discussed by Elmaghraby [25, 26]. A complete and exhaustive definition of production planning in its full generality is a complex and contentious task, as discussed by Kempf et al. [46], and must remain beyond the scope of this tutorial.

In this paper we take the position that a meaningful definition of capacity must describe the production resource's ability to convert a specified mix of inputs into a specified mix of

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

outputs within a specified time frame. This gives rise to two closely related problems. The first of these, the forward problem, is to estimate the output trajectory obtained from a production resource over time, given a specified input trajectory, to a desired level of accuracy. The backward problem is that of determining the input pattern required to produce a desired output pattern over time. To simplify our discussion we shall treat the external demand as a deterministic quantity that is known in advance. This corresponds to the common practice of planning production activities based on available demand forecasts. However, we shall assume that the actual behavior of the production resources is stochastic in nature, due to causes such as yield loss and unplanned machine breakdowns as well as natural variation in task processing times and material flows within the system. Both the forward and the backward problems as stated treat the capacity of the system as a functional in the manner suggested by Hackman [31], with its domain being the set of all possible input trajectories over time, and its range the set of all possible output trajectories over time the system is capable of producing.

The forward problem has been approached in several different ways in the literature. Clearly very simple models are possible, such as assuming that any input will converted into output after a fixed time delay, which has been widely assumed in the literature, as we shall see later in this tutorial. Another approach is to develop a set of linear or nonlinear equations representing the postulated behavior of the production system and solve these for values of the output variables given the input quantities. A special case of this latter approach is the linear programming (LP) models that are widely used for the backward problem of production planning, and are discussed below. One of the most widely used approaches is that of discrete-event simulation (Ankenman et al. [4]), in which a detailed simulation model of the production resource or facility is built and validated. This model is then loaded with a set of inputs (i.e., raw material releases) over time, and the output pattern over time predicted by the simulation model is used as an estimate of the production system's output. Another closely related approach is the use of detailed scheduling algorithms that develop detailed schedules for the loading of production resources over time, and thus describe their output.

Many deterministic scheduling models such as the widely used dispatching rules can in fact be viewed as deterministic simulations where the dispatching rule describes an aspect of the control logic governing the occurrence of discrete events such as the start and finish of job processing on each resource, as in Lu et al. [54] and Vepsalainen and Morton [72]. Other authors (Dauzère-Pérès and Lasserre [23], Pritsker and Snyder [62]) have used more elaborate scheduling algorithms that try to optimize specified performance measures such as average cycle time or resource utilization in loading the production resources. In all of these approaches, however, a set of tasks is specified together with their resource requirements and some logic describing how resource conflicts are to be resolved. The result is a pattern of output over time, together with a variety of statistics that can be computed on this output pattern. The degree of accuracy with which the simulation or scheduling model represents the actual system can, in principle, be made arbitrarily small by enhancing the level of detail captured, but generally at the cost of significant increases in computational requirements, especially when bearing in mind that multiple replications of a stochastic simulation model will be required to obtain statistically accurate estimates of the output quantities of interest.

Although queueing models (Buzacott and Shanthikumar [16], Chao et al. [20], Hopp and Spearman [36]) would appear to be a viable tool for addressing the forward problem, this has proven to be problematic. Most queueing models focus on obtaining estimates of performance measures such as the average time in system or queue length under long-run steady-state conditions, which is of limited value in the production planning context where planning periods may not be long enough to justify steady-state assumptions. Transient queueing problems, on the other hand, are notoriously difficult to analyze, but many of the

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

solution methods such as fluid models yield approaches related to the transport equation-based models addressed in §5.

The primary difficulty with the forward approach is that of the computational burden required to obtain a sufficiently accurate estimate of the output trajectory, because both simulation and scheduling models require increasing amounts of computational time as the number of resources and the number of different tasks to be scheduled increase. The forward problem is also descriptive in nature; given a specified input pattern, it describes a predicted output pattern and permits the calculation of estimated performance measures based on this pattern. The given input trajectory and the computed output trajectories may also allow computation of other quantities such as inventory levels over time.

Production planning in its classical form, as first formulated by Modigliani and Hohn [60], addresses the backward problem referred to above, which is inherently prescriptive in nature. This approach wishes to determine an input pattern to the production resource or system over time that will produce a pattern of output over time that meets external demand at minimum cost. In this case the formulation specifies the quantity of each type of output demanded by the market and attempts to compute the pattern of input over time that will meet this demand in some optimal or near-optimal manner. The results of this model specify the amount of each different type of input to be released into the production facility over time, together with the estimated output pattern produced by this input pattern.

The production planning function has as its aim the coordination of activities between different parts of the organization and external partners such as vendors and customers over time. In industrial practice this activity is generally carried out at discrete points in time, considering available information and building a plan of activities, specifically the release quantities, for a number of periods into the future, referred to as the planning horizon. This approach, which has been the dominant paradigm in the academic literature for decades (Johnson and Montgomery [40], Vollmann et al. [74]), renders a more aggregate approach possible, in which one is not trying to compute input and output trajectories in continuous time, but the total amounts of each associated with each of the discrete time periods in the planning horizon. This permits the use of simplified, aggregate constraints to represent the behavior of the production resources, which simplifies the resulting optimization models. However, there is now the possibility that the aggregate solution produced for each time period may turn out to be impossible to implement on the shop floor due to detailed resource constraints that have been ignored in the aggregation process. These problems have been addressed extensively in the literature over the last six decades using techniques of mathematical programming (Johnson and Montgomery [40], Voß and Woodruff [75]), artificial intelligence (Zweben and Fox [78]), and inventory theory (Zipkin [77]). Our focus in the first sections of this paper will be on mathematical programming models, specifically LP models of the type that have formed the basis for most academic and industrial work on these problems.

The discrete-time nature of the classical production planning formulation described above immediately raises the issues of differing time scales. The planning activity takes place periodically, using time periods usually of days or weeks; however, the actual factory for which the plans are generated operate essentially in continuous time. Thus, for a planning model to be able to produce even moderately accurate estimates of the input pattern required to produce a desired output pattern while remaining computationally tractable requires that it incorporate some model of how decisions made at the planning level will affect performance at the factory level during the planning horizon. Schneeweiss [64] refers to this type of model as an anticipation function.

Both theory and industrial practice suggest that the development of effective anticipation functions for even simple production systems is a complex task. A critical quantity for this purpose is the cycle time, the time elapsing between work being released into the production system and its emergence as finished products that can be used to meet

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

demand. The extensive literature on queueing models of manufacturing systems (Buzacott and Shanthikumar [16], Hopp and Spearman [36]) shows that under steady-state conditions the expected cycle time is a nonlinear function of the resource utilization, which, in turn, is determined by the input trajectory determined by the planning models. When we take into account that production planning rarely treats systems that can be assumed to be in steady state, and consider the number of resources and tasks involved in planning even a moderately large production system, the fundamental nature of the challenges becomes apparent. The cycle time of any item moving through the production system is necessarily a random variable subject to some probability distribution that is a function of the resource utilization, the release trajectory up to that point in time, and the resources available to perform the production tasks, among other possible factors. We shall use the term lead time to denote the estimates of cycle time (usually, its expected value) used in planning models.

In this paper we shall address a specific aspect of this problem that examines the use of discrete-event or continuous-time simulation models as anticipation functions for production planning models. We begin by describing the most common anticipation function used in optimization models of production planning, which is a fixed lead time independent of resource utilization. After presenting this model and its weaknesses, we discuss an intuitively attractive multimodel approach, which is to use a detailed simulation or scheduling model as anticipation function in an iterative scheme. We then discuss the relatively recent concept of the clearing function, which is a nonlinear anticipation function relating expected output in a planning period to some measure of expected workload during the period. Finally, we review the derivation and application to production systems of transport equation models that use systems of coupled partial differential equations (PDEs) to estimate the output pattern of a production system from a given input pattern in continuous time. We conclude this tutorial with a discussion of future research directions.

2. LP Models with Fixed Exogenous Lead Times

For simplicity, we shall begin with production planning models for a single resource producing a single product. Time will be divided into discrete time periods $t = 1, \ldots, T$, not necessarily of equal length. We wish to determine the amount of material R_t to be released to the resource during period t to meet a deterministic, known demand of D_t units in that period.

The simplest type of anticipation function arises when the cycle time of the production resources being planned are very short relative to the length of the planning periods. In this case, any material released for processing in period t is assumed to become available for use by the end of this period. If we denote the amount of material released during period t by R_t, the demand to be met during this period by D_t, and the amount of finished goods available in inventory at the end of period t by I_t, the material balance equation describing the flows of material in and out of the finished goods inventory is given by

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

In this case there is no need for the usual X_t variables denoting the amount of production in period t, because $X_t = R_t$ by assumption. Because of the very short cycle time, material remains in the system for a very short time after being released, and so considerations of work in process (WIP) can be ignored. When the cycle time of the production resource may extend over more than one planning period, a common assumption is that of a fixed lead time that is independent of resource utilization, i.e., of the release quantities R_t. This approach is commonly encountered in both inventory theory (Zipkin [77]) and mathematical programming approaches to production planning (Johnson and Montgomery [40], Voß and Woodruff [75], Missbauer and Uzsoy [59]), as well as the material requirements planning (MRP) approaches widely used in industry (Vollmann et al. [74]). The stochastic equivalent

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

of this model is a random lead time with a time-stationary probability distribution that is independent of the order quantities, such as the case treated by Eppen and Martin [27] or the class of models discussed by Zipkin [77, Chapter 7]. In this case, the amount R_t released into the system during period t becomes available for use in period $t+L$. If we denote the output of the production system in period t by X_t, we have the relationship $X_t = R_{t-L}$. The system dynamics are now described by the relationship

$$I_t = I_{t-1} + X_t - D_t = I_{t-1} + R_{t-L} - D_t. \tag{2}$$

It is common, both in the literature and in practice, to assume that the fixed lead time L corresponds to an integer number of planning periods; Hackman and Leachman [32] present a straightforward method for representing lead times that correspond to a fractional number of planning periods, which we will discuss later. As stated so far, this model assumes that there is no limit on the amount of output the system can produce in the given lead time. Most optimization models of capacitated production systems will limit the total output of the system in a given period by imposing an aggregate capacity constraint of the form $X_t \le C_t$, where C_t denotes the maximum possible output of the production system in a given period. For exposition, let us assume that each unit produced requires one time unit of the resource, and the resource capacity is expressed in terms of time units available per planning period. Under this representation, the production resource can produce any amount of output up to C_t units in any period, and no material remains in the system for more than L time periods, regardless of the work-in-process level. Thus at the end of period t, the production system in this model will have a WIP level of

$$W_t = \sum_{n=t-L+1}^{t} R_n - \sum_{n=t+1}^{t+L} X_n \tag{3}$$

units of product. It is interesting to note that this quantity does not generally appear in the constraints or objective functions of LP models of production planning, although as seen above it is not difficult to model. Note also that only R_{t-L} units of this WIP are actually available for the resource to convert into output in period t. The issue of multiple time scales raises its head once again when we examine the aggregate capacity constraint described above, because the timing of events on the shop floor is lost in the aggregation. Suppose, for instance, that the production planning model recommends that $R_t = 200$ units be released into the system during period t, where the length of the time period is one week. If the capacity of the resource is $C_t = 250$ units per week, this will appear to be a feasible solution. However, if events on the shop floor result in all 250 units arriving in the middle of the week, it may well not be possible for the resource to process all this material within period t. Subject to these limitations, we can thus describe the behavior of the production resource in a given period t with the set of constraints

$$X_t = R_{t-L}, \tag{4}$$

$$I_t = I_{t-1} + X_t - D_t, \tag{5}$$

$$X_t \le C_t. \tag{6}$$

Under the assumptions of this model, a unit of material released in period t remains in the system during the time interval defined by the periods $(t, t+L)$. To reconcile the capacity constraint with the system dynamics constraint, we must specify at what point in time during this interval the material actually occupies the capacity of the resource. Three logical constructions can be distinguished here:

(1) *lag before* models that assume that the resource capacity is occupied at the end of the lead time of this operation,

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

(2) *lag after* models where the resource capacity is occupied at the beginning of the lead time, and

(3) models that allow capacity consumption at any time within the lead time, not necessarily in a contiguous period.

For *lag before* and *lag after* models, we refer to Hackman and Leachman [32]; models of type (3) are formulated in de Kok and Fransoo [24] and Spitter et al. [68]. In this tutorial we focus on the lag before models that assume that the releases R_t in period t occupy the resource capacity in the period that the output is produced, implying $X_t = \min\{R_{t-\tau}, C_t\}$.

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

$$\sum_{n=t-\tau+2}^{t} R_n, \tag{7}$$

has no effect on the cycle time of the resource, which is always equal to the prespecified parameter τ and, as far as this model of production capacity goes, is completely unrelated to the capacity C_t of the resource in a given period. All the lead time L accomplishes is to delay the arrival of work at the resource after its release into the system; it does not describe the behavior of the resource itself, which is assumed in the capacity constraint to be able to process any amount of material up to the capacity limit C_t in a given period.

This explains an interesting anomaly with this type of model, that positive dual prices for capacity constraints result only when the capacity is fully utilized (Kefeli et al. [45]). However, queueing models repeatedly show that system performance, especially as related to WIP levels and cycle times, often begins to degrade at utilizations substantially below 1, implying the existence of situations where even though a resource is not fully utilized, additional capacity at the resource might be beneficial to system performance, although adding that capacity would not necessarily be economically desirable.

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

$$\min \sum_{t=1}^{T} (h_t I_t + c_t R_t) \tag{8}$$

subject to

$$\begin{aligned} I_t &= I_{t-1} + R_{t-\tau} - D_t && \text{for } t = 1, \ldots, T, \\ R_{t-\tau} &\le C_t && \text{for } t = 1, \ldots, T, \\ R_t &\ge 0 && \text{for } t = 1, \ldots, T, \\ I_t &\ge 0 && \text{for } t = 1, \ldots, T, \end{aligned} \tag{9}$$

where h_t is the per unit inventory cost, and c_t is the per unit production cost.

As pointed out by Hackman and Leachman [32], most LP models encountered in practice will involve additional constraints specific to the application domain under study, but the model above represents the essentials of inventory balance between periods and aggregate capacity within periods. Note that because all rates are uniformly distributed over a planning period, ensuring nonnegative inventory levels at the boundaries between periods is sufficient to ensure inventory is nonnegative throughout the period. We have chosen a simple objective function, that of minimizing the sum of production and inventory holding costs over the planning horizon. Clearly far more elaborate objective functions are possible, but our emphasis in this tutorial is on the representation of production capacity and

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

system dynamics. Finally, backlogging of any form is not allowed for purposes of exposition, although it can be incorporated in the standard manner (Johnson and Montgomery [40]).

3. Multimodel Approaches Linking LP Models and Simulation

Until now we have assumed a constant lead time L over the entire planning horizon. However, the basic structure of the LP model remains unchanged (except for initialization and ending effects at the start and end of the planning horizon) if we assume a distinct lead time L_t for the releases made in each period t. Recalling that in reality the cycle time experienced by each unit of material released in a given period is a random variable with some probability distribution f_L, the most natural interpretation of this quantity is as the expected cycle time experienced by a unit of material released in period t. It is not immediately apparent how to estimate these quantities for a given set of R_t values. Note that if we replace the maximum available capacity C_t in a period with the amount C_t^* actually used in an optimal solution, the optimal release trajectory R_t^* will not be affected. Assuming cost parameters and demand are fixed, we can now recast the production planning problem as that of searching for a consistent set of lead times L_t and capacity usage estimates $\hat{C}_t$ that will yield a minimum cost release trajectory R_t^* that produces a minimum cost output pattern X_t^* satisfying constraints (9).

However, because of the difference in time scales between the planning and execution activities, it is quite possible that a release trajectory R_t satisfying (9) may not actually be possible in the continuous time operation of the shop floor, because of errors introduced by the aggregation of the constraints. Thus testing the consistency of a proposed set of R_t, L_t, and C_t values requires solution of the forward problem. We shall assume that a viable solution to the forward problem produces estimates of the output trajectory X_t for a given input trajectory R_t such that some measure of the deviation between realized output trajectory $\bar{X}_t$ and the planned output trajectory X_t is within a prespecified tolerance with high probability. If we have available such a solution mechanism for the forward problem, a natural iteration scheme suggests itself:

(1) Set $k = 1$.

(2) Obtain initial estimates of the lead times L_t^0 and capacity usages $\hat{C}_t^0$. Set $R_t^0 = X_t^0 = 0$ for all periods t.

(3) Using the parameter estimates L_t^{k-1} and $\hat{C}_t^{k-1}$, construct and solve the LP model LP(k) to obtain the release trajectory R_t^k.

(4) Submit the release trajectory R_t^k to the forward problem solver to obtain the estimated output trajectory X_t^k and updated estimates of the realized lead times L_t^k and capacity usage estimates $\hat{C}_t^k$. If a specified stopping criterion is satisfied, stop and report the release trajectory R_t^k, predicted lead times L_t^k, and predicted output trajectory X_t^k. Otherwise, set $k = k + 1$ and go to Step 2.

This approach assumes that our forward problem solver allows us to estimate the lead times L_t^k from the input trajectory R_t^k. This is clearly true of the most widely used approach for solving the forward problem, that of discrete-event simulation, where an empirical distribution of the realized cycle times of all lots can be developed. For other approaches such as the transport equation-based models discussed in §5, these quantities are not immediately available and must be aggregated from the simulation input into an estimate compatible with their use in the LP model. For the rest of this tutorial we shall assume that the forward problem will be addressed with a simulation model.

The approach outlined above is essentially a multimodel approach that iterates between the LP model used to determine an optimal input trajectory for a given set of lead times and capacities and a simulation model that estimates the cycle times that will be realized under the given input trajectory. Most research on these iterative multimodel methods (Bang and Kim [14], Byrne and Bakir [17], Byrne and Hossain [18], Hung and Leachman [38],

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

Irdem et al. [39], Kacar [41], Kim and Kim [47]) treat the discrete-event simulation model as a fully accurate representation of the factory floor being studied. In reality, of course, the simulation model may not fully represent the reality of the shop floor, introducing errors into the input trajectories obtained from the iterative scheme. In addition, accounting for the difference in time scales between the discrete-time planning model and the essentially continuous time shop-floor simulator requires a number of nontrivial decisions on how to disaggregate the period-based decision variables for implementation in the simulation model.

To illustrate the issues involved, suppose that the duration of the basic planning period is a week, and the planning model has determined to release 200 units of product. Is the entire quantity released at the start of the week, or is it released evenly at the start of each shift? Basic queueing theory (Hopp and Spearman [36]) suggests that these decisions will have different effects on the variability of the arrival process at the resource, resulting in differences in performance. Turkseven [71] describes an algorithm for addressing this issue that is implemented in the computational study (Asmundsson et al. [12]). The converse is also true: when iterating between a simulation model and an LP model, the output of the simulation model needs to be aggregated in a manner that will provide the required parameters for the linear program, and there are potentially many different ways of doing this with potentially different consequences. These decisions, which constitute the interface between production planning and shop-floor operations, have not been studied extensively in the literature and are deserving of further attention from the research community. The algorithm of Byrne and Bakir [17] follows essentially the approach above, with the difference that it assumes all work can be completed within a single planning period, and thus iterates only on the capacity estimates $\hat{C}_t^k$. Other approaches, however, notably those of Hung and Leachman [38] and Kim and Kim [47], employ a related approach where we express the output of product i in period t, denoted Y_{it}, as

$$Y_{it} = \sum_{\tau=1}^{t} R_{i\tau} w_{i\tau t}, \tag{10}$$

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

$$\sum_{i=1}^{N} a_{ij} Y_{it} \le C_t, \tag{11}$$

because the weights $w_{i\tau t}$ would reflect the ability of the resource to produce output over time. Queueing theory tells us that these weights depend on the resource utilizations, which are determined by both the WIP profile in the system and the releases that are determined by the planning model in use. However, assuming the weights $w_{i\tau t}$ are known, we can use the capacity constraint

$$\sum_{i=1}^{N} \sum_{\tau=1}^{t} R_{i\tau} w_{i\tau t} \le C_t \tag{12}$$

together with the objective function and material balance equations of the previous model. This linear program can now be interfaced with a solver for the forward problem to provide an iterative scheme very similar to that suggested above. At the kth iteration of the procedure, the input trajectory R_t^k developed by the LP model is used to compute estimates of the weights $w_{i\tau t}^k$ and the capacity usages $\hat{C}_t^k$. The execution of the resulting plan is then simulated to obtain a new set of weight estimates, the linear program is solved with the new weights, and the procedure continues until (hopefully!) convergence. It is worth noting in passing that similar iterative techniques have been used in the job shop scheduling literature

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

to develop dispatching rules that consider the remaining time until completion of a job at an intermediate stage of processing (Lu et al. [54], Vepsalainen and Morton [72, 73]).

Hung and Leachman [38] tested a version of this iterative scheme using a simulation model of a wafer fabrication facility without any updating of the capacity usages. They associated lead times with the start of each planning period and used these lead time estimates to compute estimates of the weights $w^k_{i\tau t}$. They examined the rate of convergence of the flow time estimates to the actual flow time values in the simulation and found that convergence to the correct expected flow time values can be quite rapid, but that the procedure can fail to converge in some cases which are not fully understood. Irdem et al. [39] presented an experimental study, admittedly limited in scope to a specific production environment, of the convergence of this method that suggests that for production systems at high utilization levels it can be difficult to confirm convergence. Several possible reasons for this behavior have been examined. Irdem et al. [39] compared the $w^k_{i\tau t}$ values obtained by this procedure to those realized in the simulation and found that there can be significant differences. The use of a fully deterministic simulation does not resolve the convergence problems. One possible hypothesis is that the lead times used to estimate the weights are based on those observed for individual lots that arrive at the resource close to the beginning of the planning period, resulting in highly variable estimates of weights. Another possibility is that the implementations of the procedure have not expressly addressed the possible problem raised by Hackman and Leachman [32] that when fractional lead times are used activity rates may change within a planning period, requiring the use of additional constraints.

The other method of this type that has been studied experimentally is that of Kim and Kim [47], which differs from that of Hung and Leachman [38] in two respects: the weights $w^k_{i\tau t}$ are not computed based on the observed cycle times of specific lots in the simulation, but instead are computed directly by tracking the progress of individual lots through the system in the simulation model and the use of the realized capacity usages in the iterations. Computational experiments (Irdem et al. [39], Kacar [41]) have shown that the convergence behavior of this method is qualitatively different from that of the Hung and Leachman [38] procedure; it either converges rapidly in a few iterations, or cycles between two apparently neighboring solutions. The need for a simulation model of the system being planned requires both large amounts of data to construct and validate, and also increases run time significantly. The authors discuss several ways to reduce the computational burden of the simulation by focusing on highly utilized workcenters. Hung and Hou [37] substituted an analytical flow time prediction model for the simulation in the iterative scheme and reported results that compare favorably in convergence performance with those of the scheme using the simulation model.

A rather different iterative technique was proposed by Riaño [63], where the $w_{i\tau t}$ values are estimated using a model of the transient behavior of a queueing network. Another interesting iterative technique is that of Tian et al. [70], who addressed the problem of integrating production planning with setting safety stock levels in a multiechelon production network in the presence of stochastic demand. They iterated between an LP model for production planning and a concave nonlinear program that allocates safety stocks based on an allocation of customer demand among alternative sources. This paper is the only one we are aware of to provide necessary conditions for their iterative scheme to converge, albeit under restrictive conditions.

The iterative simulation–LP approach has the advantages of being intuitive and combining two well-understood, off-the-shelf modeling technologies. However, the formal study of these approaches is at a very early stage. There is no a priori reason to believe that convergence will occur, or even the conditions under which it may be hoped for. When we consider the general problem of searching for a set of consistent parameters (weights $w_{i\tau t}$, lead time estimates L^k_t, or capacity usages $\hat{C}^k_t$) that will allow the LP model using them to determine an optimal input trajectory, there is no reason to believe that the underlying problem is convex or

well behaved in any other way. Finally, there is currently no theoretical framework for the analysis and enhancement of these procedures, which would be an important contribution with applications beyond the domain of production planning.

An interesting alternative to the iterative approaches discussed here is the direct use of simulation optimization (Fu [28]) to determine a globally optimal input trajectory for the production facility using a discrete-event simulation model to evaluate the objective function of interest. This approach has many attractive features in principle: the simulation model can be made as accurate as desired, the interaction between the LP model and the simulation is no longer needed, and the approach is, in principle, applicable to nonconvex problems. However, the obvious drawbacks of this approach are the very high computational requirements of a simulation model of a large, complex factory and the large number of iterations required for simulation-optimization methods to converge to even a local optimum. Liu et al. [53] and Kacar et al. [43] have studied procedures of this type, finding that they are capable of finding good input trajectories but require very high central processing unit (CPU) times. Another approach of this type is given by Albritton et al. [3].

4. LP Models Using Clearing Functions

Although the iterative approach using simulation or queueing outlined above is interesting and intuitive, it suffers from a number of disadvantages, particularly the need to run time-consuming simulations of the production system being planned at each iteration. The use of fixed, exogenous lead times, on the other hand, results in tractable and well-understood optimization models, but their ability to capture the nonlinear relationships between workload and cycle time in production systems governed by queueing is questionable, especially when resources are heavily utilized or when utilization can vary significantly over time. In recent years the use of nonlinear clearing functions that represent the expected output of a production resource as a function of some measure of workload, usually the amount of WIP awaiting processing, have been proposed and show considerable promise. Here we shall give a very brief overview of this approach, referring the interested reader to Missbauer and Uzsoy [59] for a more complete review.

Recall that the basic problem induced by the differing time scales is that the planning model, which uses large, discrete planning periods, requires some means of estimating the consequences of its decisions on the performance of the shop floor. Fixed lead times are clearly a crude means of accomplishing this since they ignore the well-known queueing effects. The use of simulation or queueing models to solve the forward problem for a given input trajectory, as in the iterative schemes discussed in the previous section, is time consuming and raises complex issues of how to coordinate the two very different models used. Clearly a unified optimization model that incorporates an effective anticipation function over which optimization can take place is desirable. This is precisely the goal of the models using clearing functions.

To motivate the use of a nonlinear clearing function, consider a resource that can be modeled as a $G/G/1$ queueing system in steady state. The average number in system, i.e., the expected WIP, is given by Mehdi [55] as

$$W = \frac{c_a^2 + c_s^2}{2} \frac{\rho^2}{1-\rho} + \rho, \tag{13}$$

where c_a and c_s denote the coefficients of variation of interarrival and service times, respectively, and ρ the utilization of the server. Setting $c = (c_a^2 + c_s^2)/2$ and rearranging Equation (13), we obtain a quadratic in W whose positive root yields the desired ρ value. Solving for ρ with $c > 1$ yields

$$\rho = \frac{\sqrt{(W+1)^2 + 4W(c^2-1)} - (W+1)}{2(c^2-1)}, \tag{14}$$

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

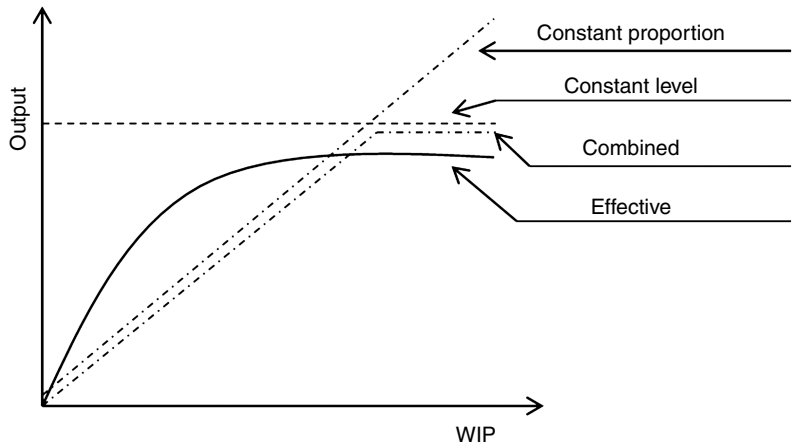
which has the desired concave form. When $0 \leq c < 1$, the other root of the quadratic will always give positive values for ρ. When $c = 1$, Equation (14) simplifies to $\rho = W/(1+W)$, again of the desired concave form. We see that for a fixed c value, utilization, and hence throughput, increase with WIP but at a declining rate due to variability in service and arrival rates. In their most general form, clearing functions represent the relationship between the expected output of a production resource in a given planning period and the some measure of the expected workload in that period. Several examples of clearing functions in the literature are depicted in Figure 1. The constant level function places a fixed upper bound on production. It does not have any lead time constraint and assumes instantaneous production no matter what the WIP level W_t is. Graves [29] proposed a clearing function in the form of $X_t = \alpha W_t$, where output X_t at time t is considered a linear function of WIP. This constant proportion function assumes a fixed lead time of $1/\alpha$ can be maintained at all utilization levels. In this model, it is assumed that production facility will be operated in the range that this fixed lead time assumption will hold. This function may yield infeasible levels of output at high WIP levels, and so needs to be limited by a fixed capacity which is shown as the combined clearing function. Karmarkar [44] proposed a nonlinear clearing function where output increases as a concave nondecreasing function of W_t, reaching an asymptotic maximum. Srinivasan et al. [69] proposed another clearing function that is a concave, nondecreasing function of WIP. Figure 1 shows these types of functions as the effective clearing function.

Missbauer [56] discussed the limitations of clearing function models such as the fact that they limit the output by a function of the expected total load and the distribution of the arrival period of the work expected to contribute to the load in period t is not considered. He proposed an aggregate order release planning model that determines the amount of work released in each planning period and can handle different load patterns without requiring additional load balancing parameters.

Early clearing function models had difficulty in modeling the behavior of multiple products because when products compete for capacity, one product may end up waiting indefinitely while other products are processed in very short lead times. To solve this problem, Asmundsson et al. [12] proposed an allocated clearing function (ACF) formulation for multiple products, where capacity at a resource is allocated to individual products.

The ACF model assumes that a clearing function will represent the aggregated output of the resource as a function of an aggregate measure of its workload, where in general the aggregation takes place over different products using the resource. An alternative approach is to use disaggregated clearing functions where a separate clearing function is estimated

Figure 1. Examples of clearing functions (Karmarkar [44]).



Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

for each product or item taking into account the workload imposed by all the different products. In this case the relationship between output and workload at the resource is represented by a set of highly coupled nonlinear constraints that are in general nonconvex (Albey et al. [2]). Current experimental evidence suggests that the ACF approach works well when products are similar in their resource consumption patterns, allowing workload to be expressed in units of processing time. However, when resource consumption patterns are more complex, such as when lot sizing or setup decisions must be made, the disaggregated approach outperforms the ACF approach.

Clearing functions can be derived analytically using steady-state or transient queueing models, or estimated empirically from empirical data. Different authors have implemented somewhat different approaches. Agnew [1] proposed a throughput function where service rate is a function of the number in queue and suggested using it in optimal control policy context. Spearman [67] derived a clearing function using closed queueing networks, conjecturing a relationship between mean cycle time and WIP, and taking one observation from simulation to specify congestion of the system. Asmundsson et al. [12] formulated the clearing function as a relationship between the expected throughput of a resource in a planning period and the time-average WIP level at the resource during the period from empirical data.

Other authors (Karmarkar [44], Missbauer [56]) assumed the clearing function depends on the expected workload, which they defined as the sum of the work in process available at the start of the period and the material released during the period. Zäpfel and Missbauer [76] used a simulation model to estimate clearing functions based on the expected workload and observed discrepancies between planned and actual WIP in simulation. Missbauer [57] showed that the clearing function depends on the work in process at the beginning of the each period due to the transient behavior of the system and suggested a transient clearing function. Selçuk et al. [65] derived transient clearing functions analytically using the Pollaczek–Khinchine mean value formula and Little's law.

The principal approach in the literature to estimating clearing functions has been to assume a concave functional form and use some form of regression to estimate the necessary parameters. Because empirical data from industrial sources are generally very difficult to obtain, most research uses a discrete-event simulation model of a production system to obtain data on the workload and output in each planning period. A notable exception is Haeussler and Missbauer [33], who fit clearing functions to data obtained from a manufacture of digital storage media. They highlighted considerable differences between then simulated and empirical data, most notably in that the empirical data show significantly higher variability that may be due to a variety of factors not included in the simulation model, such as different worker allocation policies, failures, and so on.

Asmundsson et al. [12] used a visual technique for fitting piecewise linear segments approximating a concave clearing function and reported favorable results. In a subsequent paper (Asmundsson et al. [13]) they used linear regression to fit a concave clearing function that was then piecewise linearized by solving a nonlinear program. They found that the clearing functions thus obtained consistently overestimate the capacity of resources and suggested an empirical technique to correct this by ensuring that a specified percentage of the data points lies above the fitted function. Kacar and Uzsoy [42] used different multiple regression models and found that the variable selection procedures do not seem to have significant effects on the quality of the production plans obtained by using LP models based on different fitted clearing functions. Albey et al. [2] used nonlinear regression to fit their disaggregated clearing functions, obtaining locally optimal solutions using a standard convex nonlinear solver. The majority of this work has assessed the quality of the fits obtained from the empirical data based on conventional statistical measures such as correlation coefficients. Results indicate that this area requires further research; different clearing function forms yielding very high correlation coefficients can result in quite different performance when the production plans obtained by using them in an optimization model are implemented.

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

In recent work, Kacar [41] took a different approach to the fitting of clearing functions where the parameters of the clearing function are optimized to yield the best production plans. They assessed quality of the clearing function fits by solving an optimization model using the clearing function estimated with that parameter set and then simulating the execution of the production plans in the system under study. Their approach was to start from the clearing function fit obtained by linear regression and attempt to improve this through a simulation-optimization approach, specifically the simultaneous perturbation stochastic approximation (SPSA) algorithm (Spall [66]), as well as several heuristics that attempt to reduce the required computational time. They found that significant improvements are possible over the clearing functions obtained by regression; in other words, when the clearing functions obtained by the SPSA approach are used to develop input trajectories that are then disaggregated and simulated, the clearing functions obtained by the SPSA approach have statistically better performance. In extensive computational experiments on a scaled-down semiconductor wafer fabrication system, Kacar [41] showed that the production plans using clearing functions fit by regression yield statistically better production plans than the iterative approaches of Kim and Kim [47] and Hung and Leachman [38]. However, the use of the simulation optimization to refine the parameter estimates used in the clearing functions yields substantial improvements in realized performance over the clearing functions obtained from regression alone. The additional advantage of the clearing function approach is that the actual planning does not require any simulation, although extensive simulation runs are necessary to obtain the data needed to fit the clearing functions.

5. Transport Equation Models

There has long been a tradition of aggregating the stochastic flow of products through a factory in two ways:

- averaging over time or over ensembles to convert the stochastic process into a deterministic production process and
- converting work in progress from integer variables, labeling individual parts, into a real variable describing a product continuum like a fluid.

Together these two aggregation steps convert queueing network models into ordinary differential equation (ODE) models known as *fluid models*. Queueing networks are analyzed using network equations linking the random variables describing the state of the network. *Fluid model equations* are the deterministic equations replacing the random variables with their means. Fluid models conceptually arise from treating the jobs in a queueing network as a continuous fluid that flows, via inflow and outflow rates, through a finite number of buckets (the queues),

$$\frac{dq_i}{dt} = \mu_{i-1} - \mu_i \quad i = 1, \ldots, N,$$
$$\mu_0 = \lambda,$$

where q_i is the WIP (queue-length) waiting for step i, μ_i is the processing rate of step i, and λ is the start or arrival rate into the production process. The resulting models are hybrid dynamical systems: sets of ordinary differential equations for the time evolution of the queue lengths as a function of time. The appeal of fluid models is that they are deterministic dynamical systems that are well understood even though some important issues related to the stability of queueing networks and the stability of the associated fluid models remain unresolved for multiclass queueing systems (Bramson [15], Dai [21], Dai et al. [22]).

Fluid models do not really behave like a fluid, because they still treat every production process separately and hence model the production flow through discrete steps. For long production lines with many steps, it makes sense to treat the production steps as a continuum

variable and in that way to obtain a genuine fluid dynamical description treating a factory as a pipe and parts flowing through the factory as a fluid. In contrast to a real fluid, the spatial variable does not describe physical space, but rather denotes the degree of completion of the part, or the stage of the production. Calling $x \in [0,1]$ the degree of completion, $\rho(x,t)$ describes the density of parts at stage x at time t, and $Q(t) = \int_0^1 \rho(x,t)\,dx$ is the total WIP in the factory. If the fluid moves with a velocity $v(x,t)$, then the flux (in industrial engineering terminology, the throughput) is described as $F(x,t) = v(x,t)\rho(x,t)$. Assuming that defective products are sorted out after the factory production process, there are no sources or sinks in the factory, and hence WIP satisfies the mass conservation law:

$$\frac{dQ}{dt} = \lambda - \mu, \tag{15}$$

where μ is the overall production rate of the factory. By a standard argument of transport equations (LeVeque [50]) this integral conservation law is equivalent to a differential conservation law of the form

$$\frac{\partial \rho}{\partial t} + \frac{\partial F}{\partial x} = 0. \tag{16}$$

Because $v(x,t) \geq 0$, the fluid moves from left to right. Hence we set as a boundary condition the influx $F(0,t) = \lambda(t)$; i.e., the local flux at stage zero is the arrival rate of the parts into the factory. Together with an initial WIP profile $\rho(x,0) = \rho_0(x)$, this sets up a well-defined transport equation (hyperbolic) problem.

It is instructive to consider the case of a linear wave equation with constant velocity $v(x,t) = c$. For a unit influx that is switched on at time $t = 0$ ($\lambda(t) = H(t)$, with $H(t)$ the unit step function), flowing into an empty factory, the resulting outflux generated by the solution to the transport equation (16) is

$$\mu(t) = c\rho(1,t) = H(ct-1) = \begin{cases} 0 & \text{for } t < 1/c, \\ 1 & \text{for } t > 1/c. \end{cases} \tag{17}$$

Note in particular that it takes the influx $\tau = 1/c$ time units to work its way through the factory; i.e., τ is the raw cycle time for a product.

5.1. Clearing Functions and PDEs

Using transport equations is an immediate improvement over the use of a clearing function. A delay between influx and outflux in the factory is automatically built into the model. The crucial modeling part affecting the lead time is the associated flux model:

- A constant velocity like in the previous example corresponds to a constant delay between the start and finish of an item, i.e., a constant lead time.
- A local flux at position x and time t depends only on the density around that position x. It is typically used in traffic flows and in its simplest form looks like

$$F(x,t) = v(\rho)\rho = v_0\left(1 - \frac{\rho}{R}\right)\rho. \tag{18}$$

Equation (18) is known as the Lighthill–Whitham model (Lighthill and Whitham [52]) and reflects the fact that drivers slow down as the density of cars around them increases. That process continues until the density becomes critical, $\rho_c = R$, and the velocity goes to zero, indicating a traffic jam.

- A global flux is used for a model that treats the whole factory as a single queue. As a result, the velocity at position x depends on the global quantity of total WIP, i.e., $\int_0^1 \rho(x,t)\,dx$. For instance, an $M/M/1$ queue with arrival rate λ and processing rate μ can

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

be analyzed completely in steady state. The resulting characteristics for the relationships between cycle time τ, queue length W, and influx are (Gross and Harris [30])

$$W = \frac{\lambda}{\mu - \lambda}, \tag{19}$$

$$\tau = \frac{1}{\mu}(1 + W), \tag{20}$$

$$\tau = \frac{1}{\mu - \lambda}, \tag{21}$$

$$\lambda = \frac{\mu W}{1 + W}. \tag{22}$$

The $M/M/1$ model leads to a close relationship with the clearing function idea: The cycle time is given as $\tau = 1/v$, the flux is defined as $F = v\rho$, and, identifying the queue length W in steady state with the total WIP L in the factory, we can write

$$F(\rho(x,t)) = \frac{\mu\rho(x,t)}{1 + \int_0^1 \rho(x,t)\,dx} = \frac{\mu\rho(x,t)}{1 + W(t)}, \tag{23}$$

which in steady state reduces to a clearing function of the type

$$F(W) = \frac{\mu W}{1 + W}. \tag{24}$$

Hence a PDE model combining a transport equation (16) with an $M/M/1$ flux model (23) makes the following assumptions about the production process:

- The velocity of transport, i.e., the speed at which a part moves through the production line, depends on the amount of WIP in the factory.
- The velocity at a given time is the inverse of the cycle time for an $M/M/1$ queueing system with the steady-state WIP equal to the current total load in the factory.
- The velocity is the same at every production stage, depending on the total WIP in the factory.

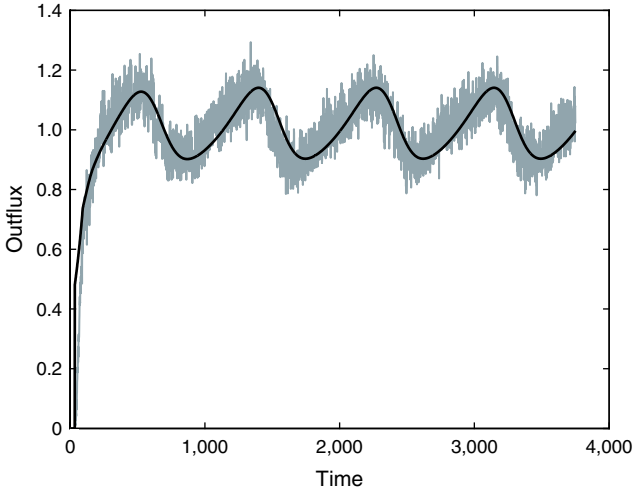
A slight generalization of this approach leads to a very usable model: By determining the state equation for the velocity $v = V(W)$, either through more elaborate queueing theory models, through measurements in the factory, or through detailed discrete-event simulations, we have a general flux model $F(x,t) = \rho(x,t)V(W(t))$ for the transport equation (16). Figure 2 (Lefeber and Armbruster [49]) compares the outflux of a discrete-event simulation with the outflux of a PDE simulation. The noisy (shaded) line comes from averaging 1,000 discrete-event simulations of a model of a semiconductor factory (Perdaen et al. [61]). The thick line is generated by solving Equation (16) with a steady-state velocity model based on relating the steady-state outflux with the corresponding steady-state WIP in the factory model.

These assumptions generate a dramatically improved model for a production process: The model captures the nonlinear dependence of the cycle times on WIP, and it correctly relates the steady-state WIP, cycle time, and outflux. In addition, it improves the outflow predictions for non-steady-state (transient) situations compared to static clearing function approaches: Variations in the outflow will be correctly resolved if they result from variations in the local product density that have little effect on the total WIP and hence on the transport velocity. Figure 2 shows that in general the outflux calculated by the PDE model stays close to the mean of the outflux for the discrete-event simulation (DES).

Notice also how the PDE model allows us to follow the transport of any local WIP portion given by $\rho(x,t)dx$ over time through the factory. Hence, if the observation time interval Δt and the cycle time τ satisfy $\tau \gg \Delta t$, a PDE model allows us to follow the flow of parts through the production unit in contrast to clearing functions or ODE-based models. On the other hand, for short cycle times, say, of the order of the observation times, the clearing

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

FIGURE 2. Comparison of a transport equation model with a discrete event simulation.



Note. Outflux generated via a sinusoidal influx for a discrete-event simulation (shaded) and a simulation based on Equation (16) and a clearing function model (black curve).

functions based on the total WIP are appropriate. This observation is independent of the velocity model that is used to describe the flow through the factory, i.e., independent of the type of clearing function that is used.

The third bullet in the above list is the most problematic feature of this model: The velocity in the factory becomes nonlocal, changing downstream when the density upstream changes. This is only a good model for highly reentrant flows with first in, first out dispatch rules. In that case, a new product entering the factory competes for production capacity with all the products already in the factory, and hence will slow down even the ones that are almost finished. At the other end of the modeling spectrum is a linear, acyclic production line, where steps downstream are completely unaware of anything that goes on upstream.

5.2. Improving the Transient PDE Model

In a series of papers, Armbruster and Ringhofer [5] and Armbruster et al. [6, 7, 9, 10] developed a kinetic theory of the stochastic transport processes in a factory model. The approach follows turbulence or gas-dynamic modeling of transport processes (Cercignani [19]). Fundamentally, such an approach is based on a probability density distribution $f(x,v,t)$, where

$$f(x,v,t)\,dx\,dv\,dt = \Pr\{\xi \in [x, x+dx], \eta \in [v, v+dv], \tau \in [t, t+dt]\}$$

describes the probability to find a particle in an x-interval with a speed in a particular v-interval in a certain time interval. A typical approach to determining the time evolution of such a probability density is to derive equations for the time evolution of its moments relative to the velocity and then to make some closure assumptions to reduce the infinite set of moment equations to a finite set (Armbruster et al. [9]). The equations for the first two moments are given by

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial v(x,t)\rho(x,t)}{\partial x} = 0, \tag{25}$$

$$\frac{\partial v(x,t)}{\partial t} + v(x,t)\frac{\partial v(x,t)}{\partial x} = 0. \tag{26}$$

Together with initial values for $\rho(x,0)$ and $v(x,0)$ and boundary conditions on the left boundary, Equations (25) and (26) are a set of well-posed hyperbolic partial differential equations.

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

Equation (26) is Burgers' equation. It models advection of a variable $v(x,t)$ that is transported along characteristics. As a result, once the initial material has left the domain, the solution is completely determined by its values at the boundary. This resolves the issue of nonlocal velocities associated with the one-dimensional model: The time it takes a part $\rho\delta x$ to move through the factory is determined at the time that the part joins the end of the queue. If there is a lot of WIP in front of the new arrival, it will take longer to clear the factory.

Whereas the boundary condition for the flux at $x = 0$ is given by the start rate $\lambda(t)$,

$$\rho(0,t)v(0,t) = \lambda(t), \tag{27}$$

the choice of the other boundary condition depends on the stochastic experiment that is described: We need to determine the expected cycle time, conditioned on the length of the queue. For an $M/M/1$ queue in steady state, the PASTA (Poisson arrivals see time averages) property suggests that an arriving part will find an average queue length W given by Equation (20). Hence,

$$v_{ss}(t) = \frac{\mu}{1+W(t)} \tag{28}$$

is the velocity related to the well-known $M/M/1$ clearing function (see Equation (24)). In the general transient case, time averages make no sense any more: We are interested to find the expected time evolution for the movement of parts through the factory given a particular initial state of the system—i.e., we are interested in the ensemble average, conditioned on the initial WIP.

Recently, Armbruster et al. [8] discussed a specific discrete-event simulation experiment and showed that we can fit heuristic boundary conditions that allow us to reproduce the ensemble averages of the experiments. The experiment was suggested by Missbauer [58]: Consider an $M/M/1$ queue with a production rate of $\mu = 1/day$, and determine the ensemble average of the cumulative production over five days as a function of the ensemble average of the total number of parts in the production unit over this five day period. Calling the initial queue $W(0)$, the ensemble average for the total number of parts becomes the sum of the initial queue and the cumulative influx $\lambda(t)$:

$$L = W(0) + \int_0^5 \lambda(t)\,dt. \tag{29}$$

Our heuristic is based on two regimes:

(1) If $\lambda(t) < \mu$, then we expect that any initial WIP distribution decays exponentially fast to the WIP distribution associated with the steady state related to the arrival rate. This is known for Markov processes as the *mixing time* (Levin et al. [51]). Hence the boundary condition is determined by the solution to an ordinary differential equation,

$$\frac{dv(0,t)}{dt} = -\sigma(v(0,t) - v_{ss}(t)) = -\sigma\left(v(0,t) - \frac{\mu}{1+W(t)}\right), \tag{30}$$

where the decay constant σ will be determined experimentally.

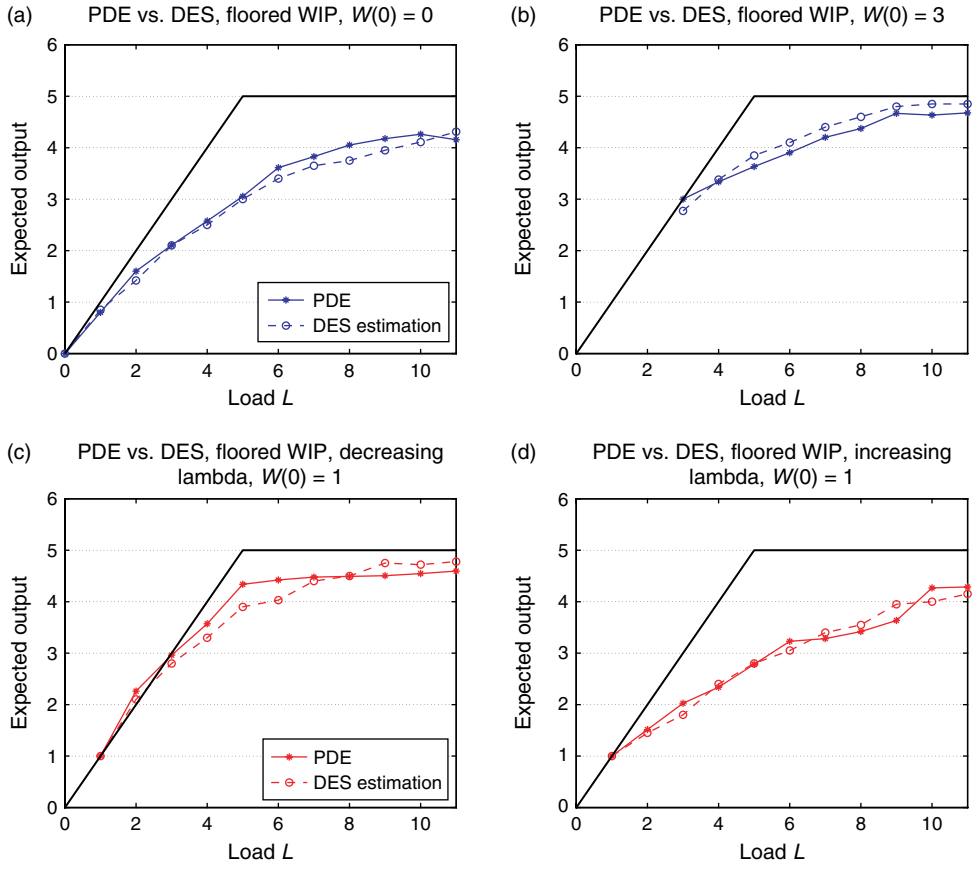
(2) If $\lambda(t) > \mu$, the queue length will become unbounded. Assuming that over the course of the observation interval of five cycle times the queue never becomes empty, and therefore the machine is never starved, the cycle time at arrival of a part at a queue length of $W(t)$ will become just $\tau = (1/\mu)W(t)$. For arrival rates only slightly larger than the production rate and for small $W(0)$, this is not always true. We found that with a velocity of

$$v(t) = \frac{\mu}{0.5+W(t)} \tag{31}$$

we obtain good agreement between PDE simulations and ensemble averaged discrete-event simulation.

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

FIGURE 3. Outflux over five time intervals as a function of the total expected load for discrete-event simulations and for the PDE model ((25) and (26)) with boundary conditions (27) and (32).



Note. This figure shows the (a) constant influx and initial WIP of $W(0) = 0$, (b) constant influx and initial WIP $W(0) = 3$, (c) decreasing influx for $W(0) = 1$, and (d) increasing influx for $W(0) = 1$.

Hence the full boundary conditions for Equation (26) become

$$\begin{aligned} v(0,t) &= \frac{\mu}{0.5 + W(t)} && \text{for } \lambda \geq \mu, \\ \frac{dv(0,t)}{dt} &= -\sigma\left(v(0,t) - \frac{\mu}{1 + W(t)}\right) && \text{for } \lambda < \mu, \\ v(0,0) &= \frac{\mu}{0.5 + W(t)}. \end{aligned} \tag{32}$$

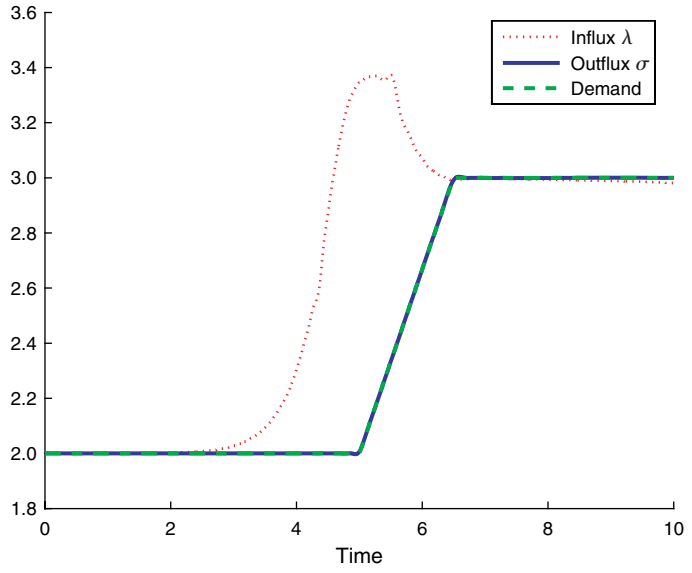
The last equation describes the initial condition for the ordinary differential equation. It is based on the assumption of a deterministic initial condition, i.e., the initial WIP is exactly known, and hence the ensemble average will be mostly affected by the stochasticity of the machine process and little affected by the stochasticity of the arrival process. Figure 3 shows comparisons of discrete-event simulations and PDE simulations of Missbauer's experiments for different influxes and different initial WIP.

5.3. Production Planning for Transport Equation Models

La Marca et al. [48] solved the production planning problem for a continuum factory model (Equation (16)) with an $M/M/1$ flux model (Equation (23)). Assuming an initial WIP distribution over the whole production line given by $\rho(x,0)$ and a target demand rate given

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

FIGURE 4. Target demand showing a linear ramp from one steady state to another (dashed line).



Note. The optimal influx (dotted line) generates the outflux (solid line) exactly on target.

for a finite time interval $d(t)$ $t \in [0, T]$, the optimal influx (input trajectory) $\lambda(t)$ is calculated that generates the least mean square error between the target demand and the outflux of the transport model. The method is based on adjoint calculus minimizing a functional of the form

$$J(\rho, \lambda) := \tfrac{1}{2} \int_o^T (d(t) - v(\rho)\rho(1,t))^2 \, dt \tag{33}$$

subject to the constraint that the outflux $F(1,t) = v\rho|_{x=1,t}$ is generated by the transport equation (16). Details can be found in La Marca et al. [48]. Figure 4 shows a typical ramp-up scenario: The optimal influx anticipates the increased outflux in by exactly the right amount at the right time to generate a production outflux that is exactly on target.

6. Conclusions and Future Directions

Despite production planning having been a focus of industrial engineering and operations research for more than five decades, the problem of representing the capabilities of even a simple production system to produce output over time remains challenging. The highly nonlinear nature of the queues representing these systems, the large size of many industrial problems, and the different sets of assumptions made by different approaches combine to render a unified, definitive resolution of this problem elusive for the foreseeable future. Despite the fact that the academic community appears to have regarded this with relatively little interest over the past two decades in contrast to, say, inventory theory, the area holds a wealth of interesting research questions with potential for real industrial and economic impact.

One need only remember the rapid success and high stock valuation of planning software firms like i2 Technologies and Manugistics in the late 1990s to get a sense of the value placed by industry on a successful solution to the problem. Linear programming models with fixed lead times are a tried and true technology with well-known theoretical limitations. A growing body of research is trying to move beyond these models and is achieving promising results in laboratory environments. However, there is still a long way to go before these tools are ready for large-scale industrial deployment.

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

The use of multimodel iterative procedures combining solutions to the forward and backward problems is as yet in its very early stages. However, the progressive development of integrated software environments such as SAS-OR, where optimization, simulation, and statistical analysis can take place in a single software system, renders the development and testing of such procedures a great deal easier than they have been in the past. The constant improvements in computing power are steadily reducing the amount of time needed to run large computational experiments with these procedures, although this is still a daunting task. Thus it is likely that we shall see more interest in these methods in the future. An important asset of these methods is their combining two well-understood modeling tools, linear programming and simulation, which have been used extensively in industry and are well understood by practitioners. However, there is very little computational experience with these methods, and not even the beginnings of a theoretical framework with which to analyze their operation and enhance their performance. Whatever method is used for solving the embedded forward problem, a clear understanding of how to link the forward and backward problems being solved to obtain reliable convergence to even a local optimum is almost completely lacking at present. The only analytical results regarding convergence of such methods are those of Tian et al. [70], which, although interesting, are very limited in scope.

The iterative multimodel approaches may be viewed as approximations to a stochastic optimization problem that searches for the optimal input trajectory treating the production system as a stochastic system. The only obvious direct way to approach this general problem at present appears to be through the use of simulation-optimization methods, whose computational requirements are prohibitive for even small problem sizes. However, pursuit of this linkage with simulation optimization may allow the formulation of an approach to study these procedures experimentally, recognize common structure between methods that may not be immediately apparent, and develop a base for analytical approaches to their performance. The principal drawback of multimodel methods involving discrete-event simulation models is their extremely high computational requirements, especially for large, complex production systems such as semiconductor wafer fabrication facilities. It is in this area that the transport equation-based methods outlined in this tutorial offer the highest promise for improvement. Extensive computational evidence shows that these techniques are capable of matching the accuracy of much more time-consuming discrete-event simulation models in very modest CPU times, immediately suggesting their use in an iterative multimodel scheme of the type we have discussed.

Optimization models using clearing functions exhibit considerable promise, but they are also at a relatively early stage of their development. A growing body of computational evidence suggests that when appropriately calibrated, these models yield production plans that outperform those obtained from multimodel methods and conventional fixed lead time models. The major research task here is the development of a sound theoretical foundation for estimating and fitting the clearing functions themselves.

Finally, it is intriguing to examine the possible relationships between the different approaches we have discussed. If we view the basic forward problem as that of modeling the behavior of a complex queueing system over time, all three approaches—clearing functions, transport equations, and multimodel approaches using simulation—are all addressing the same problem. The manner in which these different methods relate to each other may well provide important insights into how to tackle these venerable, but still challenging, problems.

The most interesting issue here is the aggregate description of transient phenomena. Although clearing functions and their implementations as flux functions in transport equations provide very good forward models and hence allow subsequent optimization schemes for the production planning problem in quasi-steady-state situations, the transient case is not even clearly defined as a problem. There are at least two different approaches:

- The state of the production facility is known exactly at a given time t_0; i.e., the WIP distribution and the conditional probabilities for machines failures in the future are known.

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

The goal is to predict the average production over a given time interval $[t_0, t-0+T]$, conditioned on the current state. This is, for instance, the issue for the recovery of a production facility after a major failure that shut down all or significant parts of the facility. In a probabilistic sense this can only be an ensemble average; i.e., we are interested in the probability distribution of the production output over this time interval as we repeat the same recovery many times.

• A related but different problem concerns the production output rate as a function of time in a time-varying environment, for instance, seasonal production or productions ramps. In that case, the initial state may not be known exactly, and we will be averaging over the ensemble of possible initial states and their influence on the subsequent production rates.

The very first attempts to generate a theory for these transient production systems follow the work of Riaño [63] for the transient behavior of queueing networks, and Selçuk et al. [65], with their analytic approximation of transient clearing functions based on stochastic processes restricted to stochastic machine processes, and Armbruster et al. [8] to generate heuristic differential equation models that model the ensemble averages for short-term production ramps. A unifying transient queueing theory for, e.g., Jackson networks seems within reach.

Acknowledgments

The research of Reha Uzsoy was supported by the National Science Foundation [Grant 1029706]. The opinions expressed in this tutorial are those of the authors and do not represent the views of the National Science Foundation. Dieter Armbruster's research was supported by a grant from the Volkswagen Foundation under the program on complex networks.

References

[1] C. Agnew. Dynamic modeling and control of congestion-prone systems. *Operations Research* 24(3):400–419, 1976.

[2] E. Albey, U. Bilge, and R. Uzsoy. An exploratory study of disaggregated clearing functions for multiple product single machine production environments. Working paper, Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, 2011.

[3] M. Albritton, A. Shapiro, and M. L. Spearman. Finite capacity production planning with random demand and limited information. Working paper, Department of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, 2000.

[4] B. E. Ankenman, J. M. Bekki, J. Fowler, G. T. Mackulak, B. L. Nelson, and F. Yang. Simulation in production planning: An overview with emphasis in recent developments in cycle time estimation. K. Kempf, P. Keskinocak, and R. Uzsoy, eds. *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook.* International Series in Operations Research and Management Science, Vol. 151. Springer-Verlag, New York, 565–592, 2010.

[5] D. Armbruster and C. Ringhofer. Thermalized kinetic and fluid models for reentrant supply chains. *SIAM Journal on Multiscale Modeling and Simulation* 3(4):782–800, 2005.

[6] D. Armbruster, P. Degond, and C. Ringhofer. Kinetic and fluid models for supply chains supporting policy attibutes. *Bulletin of the Institute of Mathematics, Academica Sinica* 66(3):896–920, 2006.

[7] D. Armbruster, P. Degond, and C. Ringhofer. A model for the dynamics of large queuing networks and supply chains. *SIAM Journal on Applied Mathematics* 66(3):896–920, 2006.

[8] D. Armbruster, J. Fonteijn, and M. Wienke. Modeling production planning and transient clearing functions. *Logistics Research*, Forthcoming.

[9] D. Armbruster, D. Marthaler, and C. Ringhofer. Kinetic and fluid model hierarchies for supply chains. *SIAM Journal on Multiscale Modeling and Simulation* 2(1):43–61, 2003.

[10] D. Armbruster, D. Marthaler, C. Ringhofer, K. Kempf, and T.-C. Jo. A continuum model for a re-entrant factory. *Operations Research* 54(5):933–950, 2006.

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

[11] K. J. Arrow, S. Karlin, and H. Scarf. *Studies in the Mathematical Theory of Inventory and Production*. Stanford University Press, Palo Alto, CA, 1958.

[12] J. M. Asmundsson, R. L. Rardin, and R. Uzsoy. Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. *IEEE Transactions on Semiconductor Manufacturing* 19(1):95–111, 2006.

[13] J. M. Asmundsson, R. L. Rardin, C. H. Turkseven, and R. Uzsoy. Production planning models with resources subject to congestion. *Naval Research Logistics* 56:142–157, 2009.

[14] J. Y. Bang and Y. D. Kim. Hierarchical production planning for semiconductor wafer fabrication based on linear programming and discrete-event simulation. *IEEE Transactions on Automation Science and Engineering* 7(2):326–336, 2010.

[15] M. Bramson. *Stability of Queueing Networks*. Lecture Notes in Mathematics, Vol. 1950. Springer-Verlag, New York, 1988.

[16] J. A. Buzacott and J. G. Shanthikumar. *Stochastic Models of Manufacturing Systems*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[17] M. D. Byrne and M. A. Bakir. Production planning using a hybrid simulation-analytical approach. *International Journal of Production Economics* 59:305–311, 1999.

[18] M. D. Byrne and M. M. Hossain. Production planning: An improved hybrid approach. *International Journal of Production Economics* 93–94:225–229, 2005.

[19] C. Cercignani. *The Boltzmann Equation and Its Applications*. Springer-Verlag, New York, 1988.

[20] X. Chao, M. Miyazawa, and M. Pinedo. *Queueing Networks: Customers, Signals and Product Form Solutions*. John Wiley & Sons, New York, 1999.

[21] J. Dai. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability* 5(1):49–77, 1995.

[22] J. G. Dai, J. J. Hasenbein, and J. H. Vande Vate. Stability and instability of a two-station queueing network. *Annals of Applied Probability* 14(1):326–377, 2004.

[23] S. Dauzère-Pérès and J. B. Lasserre. *An Integrated Approach in Production Planning and Scheduling*. Springer-Verlag, Berlin, 1994.

[24] A. G. de Kok and J. C. Fransoo. Planning supply chain operations: Definition and comparison of planning concepts. A. G. de Kok and S. C. Graves, eds. *OR Handbook on Supply Chain Management*. Elsevier, Amsterdam, 597–675, 2003.

[25] S. E. Elmaghraby. Manufacturing capacity and its measurement: A critical evaluation. *Computers and Operations Research* 18(17):615–627, 1991.

[26] S. E. Elmaghraby. Production capacity: Its bases, functions and measurement. K. Kempf, P. Keskinocak, and R. Uzsoy, eds. *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook*. Springer-Verlag, New York, 119–166, 2010.

[27] G. D. Eppen and R. K. Martin. Determining safety stock in the presence of stochastic lead times and demand. *Management Science* 34(11):1380–1390, 1988.

[28] M. C. Fu. Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing* 14(3):192–215, 2002.

[29] S. C. Graves. A tactical planning model for a job shop. *Operations Research* 34(4):522–533, 1986.

[30] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory*. John Wiley & Sons, New York, 1974.

[31] S. Hackman. *Production Economics*. Springer-Verlag, Berlin, 2008.

[32] S. T. Hackman and R. C. Leachman. A general framework for modeling production. *Management Science* 35(4):478–495, 1989.

[33] S. Haeussler and H. Missbauer. Empirical validation of metamodels of work centres in order release planning. Working paper, Department of Production and Logistics Management, University of Innsbruck, Innsbruck, Austria, 2012.

[34] F. W. Harris. *Operations and Cost*. Factory Management Series. A. W. Shaw, Chicago, 1915.

[35] C. C. Holt, F. Modigliani, J. F. Muth, and H. A. Simon. *Planning Production, Inventories and Work Force*. Prentice-Hall, Englewood Cliffs, NJ, 1960.

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

[36] W. J. Hopp and M. L. Spearman. *Factory Physics: Foundations of Manufacturing Management*. Irwin/McGraw-Hill, Boston, 2001.

[37] Y. F. Hung and M. C. Hou. A production planning approach based on iterations of linear programming optimization and flow time prediction. *Journal of the Chinese Institute of Industrial Engineers* 18(3):55–67, 2001.

[38] Y. F. Hung and R. C. Leachman. A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations. *IEEE Transactions on Semiconductor Manufacturing* 9(2):257–269, 1996.

[39] D. F. Irdem, N. B. Kacar, and R. Uzsoy. An exploratory analysis of two iterative linear programming-simulation approaches for production planning. *IEEE Transactions on Semiconductor Manufacturing* 23(3):442–455, 2010.

[40] L. A. Johnson and D. C. Montgomery. *Operations Research in Production Planning, Scheduling and Inventory Control*. John Wiley & Sons, New York, 1974.

[41] N. B. Kacar. Fitting clearing functions to empirical data: Simulation optimization and heuristic algorithms. Unpublished doctoral dissertation, Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, 2012.

[42] N. B. Kacar and R. Uzsoy. Estimating clearing functions from simulation data. B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, E. Yucesan, eds. *Proceedings of the 2010 Winter Simulation Conference*. IEEE, New York, 1699–1710, 2010.

[43] N. B. Kacar, D. F. Irdem, and R. Uzsoy. An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms. *IEEE Transactions on Semiconductor Manufacturing* 25(1):104–117, 2012.

[44] U. S. Karmarkar. Capacity loading and release planning in work-in-progess (WIP) and leadtimes. *Journal of Manufacturing and Operations Management* 2:105–123, 1989.

[45] A. Kefeli, R. Uzsoy, Y. Fathi, and M. Kay. Using a mathematical programming model to examine the marginal price of capacitated resources. *International Journal of Production Economics* 131(1):383–391, 2011.

[46] K. G. Kempf, P. Keskinocak, and R. Uzsoy. *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook*. Springer-Verlag, New York, 1–20, 2010.

[47] B. Kim and S. Kim. Extended model for a hybrid production planning approach. *International Journal of Production Economics* 73:165–173, 2001.

[48] M. La Marca, D. Armbruster, M. Herty, and C. Ringhofer. Control of continuum models of production systems. *IEEE Transactions on Automatic Control* 55(11):2511–2526, 2010.

[49] E. Lefeber and D. Armbruster. Aggregate modeling of manufacturing systems. K. Kempf, P. Keskinocak, and R. Uzsoy, eds. *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook*. International Series in Operations Research and Management Science, Vol. 151. Springer-Verlag, New York, 509–536, 2010.

[50] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, New York, 2002.

[51] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI, 2009.

[52] M. J. I. Lighthill and G. B. Whitham. On kinematic waves. II: A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society, Series A* 229:317–345, 1955.

[53] J. Liu, C. Li, F. Yang, H. Wang, and R. Uzsoy. Production planning for semiconductor manufacturing via simulation optimization. S. Jain, R. Creasey, J. Himmelspach, K. P. White, and M. Fu, eds. *Proceedings of the 2011 Winter Simulation Conference*. IEEE, New York, 3612–3622. 2011.

[54] S. Lu, D. Ramaswamy, and P. R. Kumar. Efficient scheduling policies to reduce mean and variance of cycle time in semiconductor plants. *IEEE Transactions on Semiconductor Manufacturing* 7(3):374–388, 1994.

[55] J. Mehdi. *Stochastic Models in Queuing Theory*. Academic Press, New York, 1991.

[56] H. Missbauer. Aggregate order release planning for time-varying demand. *International Journal of Production Research* 40(3):688–718, 2002.

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.

[57] H. Missbauer. Models of the transient behaviour of production units to optimize the aggregate material flow. *International Journal of Production Economics* 118(2):387–397, 2002.

[58] H. Missbauer. Order release planning with clearing functions: A queueing-theoretical analysis of the clearing function concept. *International Journal of Production Economics* 131:399–406, 2011.

[59] H. Missbauer and R. Uzsoy. Optimization models for production planning. K. Kempf, P. Keskinocak, and R. Uzsoy, eds. *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook*. Springer-Verlag, New York, 437–508, 2010.

[60] F. Modigliani and F. E. Hohn. Production planning over time and the nature of the expectation and planning horizon. *Econometrica* 23(1):46–66, 1955.

[61] D. Perdaen, D. Armbruster, K. Kempf, and E. Lefeber. Controlling a re-entrant manufacturing line via the push-pull point. *International Journal of Production Research* 46(16):4521–4536, 2008.

[62] A. A. B. Pritsker and K. Snyder. Production scheduling using FACTOR. A. Artiba and S. E. Elmaghraby, eds. *The Planning and Scheduling of Production Systems.* Chapman & Hall, Boston, 337–358, 1997.

[63] G. Riaño. Transient behavior of stochastic networks: Application to production planning with load-dependent lead times. Unpublished doctoral thesis, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, 2003.

[64] C. Schneeweiss. *Distributed Decision Making*. Springer-Verlag, Berlin, 2003.

[65] B. Selçuk, J. C. Fransoo, and A. G. de Kok. Work in process clearing in supply chain operations planning. *IIE Transactions* 40:206–220, 2007.

[66] J. C. Spall. Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Transactions on Aerospace and Electronic Systems* 34(3):817–823, 1998.

[67] M. L. Spearman. An analytic congestion model for closed production systems with ifr processing times. *Management Science* 37(8):1015–1029, 1991.

[68] J. M. Spitter, C. A. J. Hurkens, A. G. de Kok, J. K. Lenstra, and E. G. Negenman. Linear programming models with planned lead times for supply chain operations planning. *European Journal of Operational Research* 163:706–720, 2005.

[69] A. Srinivasan, M. Carey, and T. E. Morton. Resource pricing and aggregate scheduling in manufacturing systems. Working paper, Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, 1988.

[70] F. Tian, S. P. Willems, and K. G. Kempf. An iterative approach to item-level production and inventory planning. *International Journal of Production Economics* 133(1):439–450, 2011.

[71] C. H. Turkseven. Computational evaluation of production planning formulations using clearing functions. M.S. thesis, School of Industrial Engineering, Purdue University, West Lafayette, IN, 2005.

[72] A. P. Vepsalainen and T. E. Morton. Priority rules for job shops with weighted tardiness costs. *Management Science* 33(8):1035–1047, 1987.

[73] A. P. Vepsalainen and T. E. Morton. Improving local priority rules with global lead-time estimates: A simulation study. *Journal of Manufacturing and Operations Management* 1:102–118, 1988.

[74] T. E. Vollmann, L. Berry, D. C. Whybark, and F. R. Jacobs. *Manufacturing Planning and Control for Supply Chain Management*. McGraw-Hill, New York, 2005.

[75] S. Voß and D. L. Woodruff. *Introduction to Computational Optimization Models for Production Planning in a Supply Chain*. Springer-Verlag, Berlin, 2003.

[76] G. Zäpfel and H. Missbauer. Production planning and control (PPC) systems including load-oriented order release—Problems and research perspectives. *International Journal of Production Economics* 30:107–122, 1993.

[77] P. H. Zipkin. *Foundations of Inventory Management*. Irwin, Burr Ridge, IL, 2000.

[78] M. Zweben and M. Fox. *Intelligent Scheduling Systems*. Morgan Kaufmann, San Francisco, 1994.

Downloaded from informs.org by [216.73.216.247] on 14 November 2026, at 09:25 . For personal use only, all rights reserved.