Chapter 3
Integrated Production Planning and Pricing Decisions in Congestion-Prone Capacitated Production Systems

Abhijit Upasani and Reha Uzsoy

Introduction

The highly capital intensive nature of the semiconductor industry requires its factories to operate at high utilization levels, where small changes in utilization can cause large changes in lead times. Demand for these products over time can be quite uneven, which leads to firms trying to shape their demand by price promotions in order to maintain high factory utilization levels. However, it is well known from queuing models of manufacturing systems (Buzacott and Shanthikumar 1993) that higher resource utilization leads to increasing lead times. This raises the possibility of price reductions becoming counterproductive—an unwise price promotion can create a surge in demand that, after some time, results in a large increase in lead times, missed delivery dates, cancelled orders and lost future business.

To this end, companies will often develop aggregate production plans at the product family level for several months (up to 18 months in the case of one semiconductor manufacturing firm described by Allison et al. (1997)) in order to identify potential capacity bottlenecks and make sure that competitive lead times can be maintained. This plan, based on current order books and marketing forecasts, permits the planning of price promotions as part of the process. Given the long planning horizon, an aggregate planning model focusing on the loading of resources and management of prices over time to achieve maximum profit with competitive lead times would be useful to management. The high utilization levels at which many capital-intensive factories, such as semiconductor wafer fabs, operate renders a planning model that accounts for the nonlinear relationship between resource utilization and lead times desirable, especially when customers are sensitive to both lead times and prices.

R. Uzsoy (✉)
Edward P. Fitts Department of Industrial and Systems Engineering,
300 Daniels Hall, Campus Box 7906, North Carolina State University,
Raleigh, NC 27695–7906, USA
e-mail: ruzsoy@ncsu.edu

A. Upasani
Terra Technology, 20 Glover Avenue, Norwalk, CT 06850, USA
e-mail: abhijit.upasani@gmail.com

P. S. Pulat et al. (eds.), *Essays in Production, Project Planning and Scheduling,*
International Series in Operations Research & Management Science 200,
DOI 10.1007/978-1-4614-9056-2_3, © Springer Science+Business Media New York 2014

Most existing pricing-production planning models do not address this problem in its full complexity. In particular, most such models do not consider the effect of workload on queues and lead times, and hence may underestimate the price that should be charged at a given output level. In particular, if high demand results in long lead times due to congestion in the production facility, the assumption that demand can be met within a fixed lead time may result in significant lost sales. Dynamic pricing models based on queuing, on the other hand, generally describe long-run steady-state behavior and do not provide a framework for decisions to be made over time.

The model presented in this chapter is a first step towards addressing these issues. We use clearing functions (CFs, Asmundsson et al. 2009) to capture the nonlinear relationship between resource utilization and lead times. Following the literature, customer behavior is modeled using a demand function that is linear in both price and lead time, with a maximum lead time beyond which no demand will be forthcoming. In each planning period, customers can observe the average flow time associated with the current workload of the production system, and place orders accordingly. Such systems are already in use by semiconductor manufacturing companies such as Taiwan Semiconductor Manufacturing Corporation which provide contract manufacturing services to other firms (www.tsmc.com/english/dedicatedFoundry/services/eFoundry.htm). The model jointly determines the price and the amount of work to be released in each time period, thus determining the average lead time associated with that planning period. The model allows the possibility of production smoothing through the accumulation of finished goods inventories and price promotions.

Our results show that when the demand is sensitive to lead times, the CF model with workload-dependent lead times produces significantly higher profits than a conventional model assuming a fixed lead time. In several scenarios the release plans suggested by the fixed lead time model are unable to satisfy the market demand generated by the associated prices, since they assume that a fixed lead time can be maintained in the face of the high demand created by low prices. In fact, the increased demand resulting from price reductions can only be met with long lead times, which end up reducing demand. Hence a thorough understanding of the effects of pricing on lead times and queues is essential for capacity constrained firms that plan to use dynamic pricing. As suggested by Pekgun et al. (2008), the separation of lead time and pricing considerations between the production and marketing operations is a significant obstacle to this understanding, suggesting the need for more emphasis on this interface in capital-intensive firms operating at high utilization levels.

Literature Review

Our research is related to three different streams of literature: joint pricing and production planning models, models for load-dependent lead-time quotation, and steady state models that study relationships between price and lead times.

Joint pricing and production planning models aim to produce a profit-maximizing combination of prices and production plans. Eliashberg et al. (1991) and Yano and

Gilbert (2003) present detailed reviews of this stream of literature. This literature also includes dynamic pricing models that change prices over time to improve profitability (Swann 2001; Charnsirisakskul et al. 2006; Deng and Yano 2006). Ahn et al. (2007) present an interesting model where demand in a given period depends on prices in preceding periods. Adida and Perakis (2006) consider a continuous time model with a linear demand function and an additive model of uncertainty, and present a robust optimization model. In a subsequent paper (Adida and Perakis 2010) they compare robust and stochastic optimization models for this problem, noting that stochastic optimization models can be sensitive to the probability distributions used. The related area of dynamic pricing focusing on the interface with inventory management is reviewed by Elmaghraby and Keskinocak (2003).

Researchers in this area have used simple, aggregate capacity constraints with limited ability to consider interactions between capacity utilization and lead time. When faced with high demands that saturate capacity constraints in a given period, these models will build inventory in earlier periods, effectively increasing lead times. However, this dynamic does not capture the rapid nonlinear increase in lead times observed at higher utilizations, providing an incomplete picture of system behavior. Recent work (Kefeli et al. 2011) has shown that in the presence of congestion the theoretical output of the system may not be achieved due to the very high work in process inventories required to achieve them, causing these types of capacity constraints to give an optimistic picture of the production system's ability to meet demand. We illustrate this effect in our numerical examples.

Chen and Hall (2010), in contrast, consider the pricing of individual orders on a single machine or a two-machine flow shop to maximize profit under different cost criteria which are determined by the production schedules. They provide exact dynamic programming algorithms and heuristics, and demonstrate that even heuristic solutions to the problem yield significant improvement in profit over the case where prices and schedules are determined independently. Since these models represent capacity at a very fine level of detail, they capture the relationship between utilization and lead times correctly. However, such models do not easily scale up to the longer time periods addressed in this chapter.

The second stream of literature encompasses models that estimate lead times based on the current state of the system and use these lead times for order negotiation. These models recognize that lead times are load dependent and address operational decisions like input control or order selection, price and lead-time quotation, and capacity investment (Donohue 1994; Easton and Moodie 1999; Elhafsi and Rolland 1999; Elhafsi 2000; Charnsirisakskul et al. 2004; Plambeck 2004). While these models allow marketing to make realistic lead-time quotations to be used in price negotiation, they do not capture the relationship between prices, resource utilization and lead times.

A related set of models, classified in the literature as order acceptance models, assume stochastic (usually Poisson) customer arrivals and quote each customer a delivery date based on system status (Dellaert 1991; Duenyas 1995; Duenyas and Hopp 1995). These models assume a certain probability that the customer will actually place an order when quoted the delivery date, thus obtaining an effective arrival rate for orders. Late orders are penalized and the models aim to minimize the impact of this penalty on revenue, which is fixed for every order.

The last stream of models conducts steady state analyses of relationships between price, lead time and capacity for *M/M/*1 systems (Low 1974; Palaka et al. 1998; So and Song 1998; Boyaci and Ray 2003; Ray and Jewkes 2004). Almost all these models use a demand function that is linear in both price and lead time to represent the market and aim to set prices and lead times subject to a service level constraint under steady state conditions. These models yield useful managerial insights through their characterization of optimal behavior, but their steady state nature does not allow them to be used to develop pricing and production plans over a finite horizon. Liu et al. (2007) study price and lead-time setting in a decentralized supply chain where a supplier specified a wholesale price and a planned delivery time, while the retailer quotes a retail price. Customers are sensitive to both lead time and retail price. They model the behavior of the supplier and retailer as a Stackelberg game and obtain the equilibrium strategy of both actors. Pekgun et al. (2008) developed a steady-state make-to-order (MTO) model that incorporates coordination mechanisms for price and lead-time quotation.

Plambeck (2004) considers capacity setting, price and lead-time quotation, and order sequencing decisions in a MTO system with two customer classes and compares dynamic against static lead-time quotations (similar to our Fixed Lead Time (FLT) model). The key assumption the author makes is that customers belonging to the "patient" class will tolerate long lead times. The author requires this slow-moving portion of the order queue to be so large that the system utilization approaches 100 %, allowing the author to apply heavy-traffic queuing approximations to derive optimal decision policies. Our CF model considers a different problem, that of determining an integrated aggregate plan for factory loading and pricing over discrete time periods in the face of the market's sensitivity to lead times. Our model does not impose a utilization level on the system but instead allows the system to choose its optimal utilization level. Consistent with Plambeck's results, our model also shows that taking the state of the system into account can yield significantly higher profit than a fixed lead-time model.

The joint planning models in the first stream represent aggregate planning decisions in a make-to-stock environment, where a different price is quoted every period, but all orders in the same period observe the same price. These joint planning models fall under the domain of models at the production/marketing interface that also includes models for sales-production coordination mechanisms (Eliashberg and Steinberg 1991; Upasani and Uzsoy 2008). Models in the last two environments focus on a MTO environment where no stocks of finished goods inventory are held and each order can be quoted a separate price or lead time. Detailed reviews of models in the last two streams are found in Chatterjee et al. (2002), Keskinocak and Tayur (2004), and Upasani and Uzsoy (2008).

To summarize the existing literature, joint planning-pricing models have limited ability to capture the effects of utilization on delivery times, whereas steady-state lead-time quotation models do not yield medium-term plans over a finite horizon. Recent developments in production planning models with load-dependent lead times

(Pahl et al. 2005) provide avenues for integrating state-dependent lead times into models of the production-marketing interface. Specifically, we use CFs (Pahl et al. 2005, 2007; Asmundsson et al. 2006; Asmundsson et al. 2009; Missbauer and Uzsoy 2010), which relate the expected throughput of a production system in a planning period to the expected work-in-process (WIP) inventory level over the period.

Clearing Functions

A promising approach to modeling workload-dependent lead times in production planning has been the use of CFs (Karmarkar 1989) that represent the expected output of a resource over a given period of time as a function of the expected WIP inventory level over that period. The term has its origin in work by Graves (1986) that specifies the fraction of the current WIP that can be processed to completion ("cleared") by a resource in a given time period. Karmarkar (1989) and Srinivasan et al. (1988) independently develop nonlinear CFs for production planning models. We shall use the term "WIP" to denote any reasonable measure of the WIP inventory level over a period of time that can be used as a basis for a CF. An extensive review of CFs and their use in production planning models is given by Missbauer and Uzsoy (2010)

To motivate the use of a nonlinear CF, consider a resource that can be modeled as a *G/G/*1 queuing system in steady state. The average number in system, i.e., the expected WIP, is given by Medhi (1991) as

$$w = \frac{(c_a^2 + c_s^2)}{2}\frac{\rho^2}{(1-\rho)} + \rho \tag{3.1}$$

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 $c = (c^2{}_a + c^2{}_s)/2$ and rearranging (1), we obtain a quadratic in W whose positive root yields the desired ρ value. Solving for ρ with $c > 1$ yields

$$\rho = \frac{\sqrt{(W+1)^2 + 4W(c^2-1)} - (W+1)}{2(c^2-1)} \tag{3.2}$$

which has the desired concave form. When $0 \le c < 1$, the other root of the quadratic will always give positive values for ρ. When $c = 1$, (3.2) 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.

Several authors discuss the relationship between throughput and WIP levels in the context of queuing analysis, where the quantities studied are the long-run steady-state expected throughput and WIP levels. Agnew (1976) studies this type of behavior in the context of optimal control policies. Spearman (1991) presents an analytic congestion model for closed production systems with increasing failure rate processing time distributions that describes the relationship between throughput and WIP. Hopp

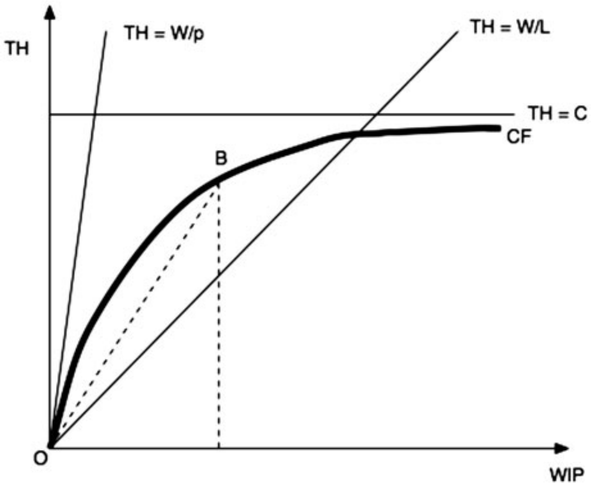

Fig. 3.1 Examples of clearing functions (Karmarkar 1989)

and Spearman (2001) provide a number of illustrations of CFs for a variety of systems. Srinivasan et al. (1988) derive the CF for a closed queuing network with a product form solution. While these approaches are based, as is our analysis above, on steady-state queuing models, a number of researchers have examined the issue of estimating CFs when the underlying queuing system is not in steady state. Asmundsson et al. (2009) show that even under transient conditions the concave shape of the CF will be maintained. Missbauer (2009) and Selçuk (2007) use transient queuing models to derive CFs under somewhat different sets of assumptions.

Figure 3.1, derived from Karmarkar (1989) depicts several examples of CFs considered in the literature to date. The horizontal line $TH = C$ corresponds to a fixed upper bound on output over the period, but without a lead-time constraint it implies that production can occur without any WIP in the system if work release and production are synchronized. This is reflected in the independence of output from the WIP level, which may constrain throughput to a level below the upper bound by starving the resource. This approach is implemented in, for example, the Capacitated MRP (MRP-C) approach of Tardif and Spearman (1997) and most LP approaches such as that of Hackman and Leachman (1989), but is supplemented with a fixed lead time that is an exogenous parameter independent of workload. The linear CF of Graves (1986) is represented by the $TH = W/L$ line, which implies a lead time of L periods that can be maintained independently of the WIP level. Note that if WIP and output are measured in the same time units (e.g., hours of work), the slope of the proportional part of the function is $1/L$, where L is the average lead time. However, as seen in Fig. 3.1, this model may suggest infeasible output levels when WIP levels are high. If a fixed lead time is maintained up to a certain maximum output, we have the relationship $TH = min\{W/L, C\}$. In the special case of the Graves CF where the lead time is equal to the average processing time, with no queuing delays at all, we obtain the line $TH = W/p$, where p denotes the average processing time. Assuming that average lead time is equal to the average processing time up to the maximum output level, it gives the "Best Case" model $TH = min\{W/p, C\}$ described in Chap. 7 of Hopp and Spearman (2001). However, by linking production rate to WIP level, a linear CF differs from the fixed delays used in most LP models, where the output of a production process is simply the input shifted forward in time by the fixed lead time. Orcun et al. (2006) illustrate the differences between these models using system dynamics simulations. For most capacitated production resources subject to congestion, limited capacity leads to a saturating (concave) CF. It is also apparent from the Fig. 3.1 that the CF always lies below the $TH = W/p$ and $TH = C$ lines.

An important issue in using CF models is the question of how long the planning periods should be. If the CFs are derived using steady-state queuing models, the planning period must be long enough that the queues representing the production system can be at least approximately in steady state. Given the long-term, aggregate purpose of this type of model, as outlined in the introduction and the discussion in Allison et al. (1997), the planning buckets are likely to be long enough (e.g., a month) that most production systems with relatively short raw processing times should reach steady state. However, even if this is not the case, current research is exploring means of deriving CFs for systems under transient regimes (for example, Selcuk 2007 and Missbauer 2009), showing that even under transient conditions the concave shape of the CF is maintained.

A number of authors have suggested empirical approaches to estimating CFs, where a functional form with the desired properties is postulated, and then fit to data obtained either from an industrial facility or a simulation model using some form of regression analysis. Karmarkar (1989) suggests a CF of the form

$$X_t = \frac{K_1 W_t}{K_2 + W_t} \tag{3.3}$$

where X_t denotes the output in period t, W_t the WIP at the resource at the start of period t, and K_1 the maximum possible output of the resource in period t. The shape parameter K_2 is estimated by the user. Selçuk et al. (2007) demonstrate the derivation of K_2 for an $M/G/1$ system with bulk arrivals. Srinivasan et al. (1988) suggest an alternative functional form

$$f(W_t) = K_1(1 - e^{kW_t}) \tag{3.4}$$

where k is again a user-estimated shape parameter. Asmundsson et al. (2009) use this latter functional form and give an extensive discussion of various issues in collecting simulation data for fitting this type of CF. Asmundsson et al. (2006) use a visual fit of linear segments to simulation data to develop a CF formulation for a scaled-down semiconductor wafer fabrication facility with unreliable equipment and reentrant flows. Kacar and Uzsoy (2010) and Kacar et al. (2010) use a linear regression approach applied to data collected from a simulation model, with good results. Asmundsson et al. (2009) show that an empirically fitted CF can give good results even under a transient regime. The implication for this research is that it is possible to represent the behavior of a production system with an appropriately fitted

CF. Thus we shall proceed with our model on the assumption this can be done and examine the potential impact on profits of using a model with fixed lead time that does not consider queuing behavior.

A Single Product Dynamic Joint Price-Production Model Incorporating Congestion

We now present a joint price-production model that incorporates CFs and lead-time-dependent demand. We assume a single firm that behaves as a monopolist. The firm sees a linear demand function $D = g(P, L) = Max\{0, M\text{-}aP\text{-}bL\}$, where $a, b \geq 1$ are the price and lead-time sensitivities of demand D with respect to price P and lead time L, respectively. Changes in market conditions are represented by changes in these sensitivities. The intercept M of the demand function represents the maximum possible demand, i.e., the market size.

In a given period t, the firm quotes a price P_t and a delivery time L_t to customers. We assume that the firm quotes a delivery time for orders received in a period equal to the average manufacturing lead time at the start of the period. Since the manufacturing lead time (delivery time) depends on the number of orders waiting, the firm can control the *maximum* delivery time by limiting the number of orders to be processed (per Fig. 3.1). In effect, the firm quotes the delivery time based on the minimum of two values: the average manufacturing lead time, and a guaranteed delivery time L_G by which all orders need to be satisfied, or customers will not place orders. Hence an order received in period t has to be fulfilled by period $t + L_G$.

The firm needs to align its production system with this market preference by mapping L_G on Fig. 3.1 and quoting an average delivery time below the value of L_G. This will, in turn, determine the number of orders that a firm may accept and hold in queue for processing, yielding a target production rate and a target utilization. Thus, the higher the guaranteed delivery time allowed by the market, the higher the utilization at which the firm can operate its resources. From Fig. 3.1, as utilization increases, a large increase in threshold value L_G will allow only a small increase in utilization, since lead time increases rapidly with additional workload at high utilization levels. This guaranteed delivery time assumption is similar to that used by Selcuk et al. (2007) and Spitter et al. (2005a; b) in their supply chain operations planning (SCOP) models, where they assume a planned manufacturing lead time within which an order must finish processing. The idea of a quoted lead time in combination with a maximum lead time is also used by Dellaert (1991) and Duenyas and Hopp (1995) in their models of due-date management with order selection.

Another mechanism by which a firm may control quoted average delivery time is to quote a higher price and thus accept fewer orders. This is possible due to the monopolist assumption and the price and lead-time-dependent nature of demand. Customers may be willing to pay a premium for lower-quoted average delivery times and the relative magnitude of this premium would depend upon their sensitivity to delivery time represented by parameter b_t.

The average delivery time quotation implies that some orders will be ready for delivery earlier than promised. The customer may not always want to take delivery early, in which case the manufacturer has to hold finished goods inventory. We assume that the customer will allow a limited number of orders to be delivered early in the planning horizon and represent this by a parameter ν. Late deliveries are not allowed, though this can be incorporated in a straightforward manner. To further simplify the model, we restrict every order to have a size of one unit. This enables us to eliminate constraints that would otherwise be included to track fulfillments of orders of varying sizes. We define the following notation:

Variables

R_t Order release quantity in period t
W_t WIP inventory at the end of period t
X_t Production quantity in period t
I_t Finished goods inventory (FGI) at end of period t
P_t Price in period t
D_t Sales quantity in period t
Y_t Quantity shipped in period t

Parameters

a_t Price sensitivity of demand in period t
b_t Lead-time sensitivity of demand in period t
h_t Holding cost of finished goods inventory per unit in period t
ω_t Holding cost of WIP inventory per unit in period t
ϕ_t Unit production cost in period t
c_t Order release cost per unit released in period t
ν Maximum units allowed to be shipped before due date over the horizon
K_1 Theoretical maximum production capacity
K_2 Curvature parameter of CF
M Intercept of demand function, i.e., demand when price = lead time = 0
T Length of planning horizon, $t = 1,..,T$
L_G Guaranteed delivery time (in periods)
f(.) CF

Let $\hat{W}_t$ be the estimated average WIP level in a period t. We use the CF form suggested by Karmarkar (1989). From (3.3) we have

$$f\left(\hat{W}_t\right) = \frac{K_1 \hat{W}_t}{K_2 + \hat{W}_t}$$

The production X_t in period t is bounded by the CF in that period.

As mentioned earlier, the demand in period t is expressed by the demand function $D_t = M - a_t P_t - b_t L_t$. By Little's Law, the expected lead time in period t is given by $L_t = \hat{W}_t / X_t$, expressed in units of periods. By invoking Little's Law we assume that the production system is in steady state within the planning period. Thus, the demand observed in period t is given by

$$D_t = M - a_t P_t - b_t \left(\frac{\hat{W}_t}{X_t} \right)$$

Our CF-based joint price-production planning model is now as follows:

CF model

$$Max \sum_{t=1}^{T} \left[P_t \left(M - a_t P_t - b_t \left(\frac{\hat{W}_t}{X_t} \right) \right) - c_t R_t - \phi_t X_t - h_t I_t - \omega_t W_t \right] \tag{3.5}$$

s.t.

$$\{\lambda_t\} \quad W_t = W_{t-1} - X_t + R_t \qquad \forall t \tag{3.6}$$

$$\{\pi_t\} \quad I_t = I_{t-1} + X_t - Y_t \qquad \forall t \tag{3.7}$$

$$\{\theta_t\} \quad X_t \leq \frac{K_1 \hat{W}_t}{K_2 + \hat{W}_t} \qquad \forall t \tag{3.8}$$

$$\{\mu_t\} \quad M - a_t P_t - b_t \left(\frac{\hat{W}_t}{X_t} \right) \geq 0 \qquad \forall t \tag{3.9}$$

$$\{\sigma_t\} \quad \sum_{\tau=1}^{t} Y_\tau \geq \sum_{\tau=1}^{t-L_G} \left[M - a_\tau p_\tau - b_\tau \left(\frac{\hat{W}_\tau}{X_\tau} \right) \right] \qquad \forall t \tag{3.10}$$

$$\{\rho_t\} \quad \sum_{\tau=1}^{t} Y_\tau \geq \sum_{\tau=1}^{t-L_G} \left[M - a_\tau p_\tau - b_\tau \left(\frac{\hat{W}_\tau}{X_\tau} \right) \right] + \nu \qquad \forall t \tag{3.11}$$

$$\{\chi_t\} \quad \hat{W}_t \leq \frac{1}{2}(W_{t-1} + W_t) \qquad \forall t \tag{3.12}$$

$$P_t, X_t, I_t, W_t, R_t, Y_t, \hat{W}_t \geq 0 \qquad \forall t \tag{3.13}$$

The objective is to maximize total contribution, expressed as the difference between the total revenue in each period and variable operating costs. Equations (3.6) and (3.7) are WIP and finished goods inventory balance constraints. Equation (3.8) represents production capacity using the CF, and constraint (3.9) defines the sales quantity. Constraint (3.10) requires that all orders be shipped within the planned delivery time, but allows orders to be shipped earlier than due, rather than being held as finished goods inventory. Since the customer may impose a limit on the number of orders shipped early over the horizon (given by the parameter ν), we model this preference in constraint (3.11). We estimate the average WIP level $\hat{W}_t$ within a given period using the WIP levels at the end points of the period using (3.12). All variables are required to be non-negative by (3.13). The Greek letters in curly brackets to the left of each constraint denote its associated Lagrange multipliers. We do not impose a cost on shipping since it would require setting values for another parameter, which we avoid for sake of parsimony in the experimental design. For the same reason, we do not impose a penalty if the average delivery time quotation exceeds the planned delivery time. Instead we reduce sales through our time-dependent demand function. This mechanism is further discussed in the section “Experiments Without Early Delivery Flexibility: $\nu = 0$.

For comparison purposes we now state a joint price-production planning model that assumes a fixed delivery time $L \leq L_G$ which is specified as an exogenous parameter, and hence is denoted as the Fixed Lead Time (FLT) model. The demand observed by this model in period t is expressed as $D_t = M - a_t P_t - b_t L$. The firm must set $L \leq L_G$ to avoid exceeding the target utilization. We assign a Lagrange multiplier for each constraint as was done for the CF model.

FLT model

$$Max \sum_{t=1}^{T} \left[\hat{P}_t \left(M - a_t \hat{P} t - b_t L \right) - c_t \hat{X}_t - h_t \hat{I}_t \right] \tag{3.14}$$

s.t.

$$\{\gamma_t\} \qquad \hat{I}_t = \hat{I}_{t-1} + \hat{X}_{t-L} - \hat{Y}_t \qquad \forall t \tag{3.15}$$

$$\{\delta_t\} \qquad \hat{X}_t \leq K_1 \qquad \forall t \tag{3.16}$$

$$\{\hat{\mu}_t\} \qquad M - a_t \hat{P}_t - L \geq 0 \qquad \forall t \tag{3.17}$$

$$\{\hat{\sigma}_t\} \qquad \sum_{\tau=1}^{t} \hat{Y}_\tau \geq \sum_{\tau=1}^{t-L} (M - a_\tau \hat{P}_\tau - b_\tau L) \qquad \forall t \tag{3.18}$$

$$\{\hat{\rho}_t\} \qquad \sum_{\tau=1}^{t} \hat{Y}_\tau \geq \sum_{\tau=1}^{t-L} (M - a_\tau \hat{P}_\tau - b_\tau L) + \nu \qquad \forall t \tag{3.19}$$

$$\hat{X}_t, \hat{P}_t, \hat{Y}_t, \hat{I}_t \geq 0 \qquad \forall t \tag{3.20}$$

We use the variable $\hat{X}_t$ to denote production initiated in period t. Since there is a fixed production lead time L, production initiated in period t is available to be shipped in period $t + L$. This variable corresponds to the releases variable R_t from the CF model. Hence we incorporate a time lag L in the inventory balance constraint (3.15). Since the FLT model ignores the buildup of queues in the system due to its fixed lead-time assumption, it does not have any WIP variables or WIP balance constraints. This model is consistent with FLT production planning models (Johnson and Montgomery 1974; Hackman and Leachman 1989; Spitter et al. 2005a; Spitter et al. 2005a) or the joint price-production model of Swann (2001).

In our numerical experiments we use a modified version of the FLT model that facilitates direct comparisons with the CF model. Recall that $\hat{X}_t$ models the material released in period t so that it finishes processing and is available for shipping in period $t+L$. This definition, while capturing the nature of fixed lead times, does not allow a direct comparison between the two models. Hence, we replace the variable $\hat{X}_t$ with two variables: $\hat{R}_t$ to denote the material release in period t, and $\hat{X}'_t$, the actual production in period t. The two variables are related by the expression $\hat{R}_t = \hat{X}'_{t+L}$. Hence, the $\hat{R}_t$ units of work released in period t will remain in WIP for L time periods, which we explicitly include in the objective function. The modified FLT model is thus as follows:

$$\sum_{t=1}^{T} \left[\hat{P}_t \left(M - a_t \hat{P}_t - b_t L \right) - c_t \hat{R}_t - \phi_t \hat{X}'_t - h_t \hat{I}_t - \omega_t \sum_{j=t-L+1}^{t} \hat{R}_j \right] \tag{3.21}$$

s.t.

$$\hat{X}'_{t+L} = \hat{R}'_t \qquad \forall t$$

$$\hat{I}_t = \hat{I}_{t-1} + \hat{X}'_t - \hat{Y}_t \qquad \forall t \tag{3.22}$$

$$\hat{X}'_t \leq K_1 \qquad \forall t \tag{3.23}$$

$$M - a_t \hat{P}_t - b_t L \geq 0 \qquad \forall t \tag{3.24}$$

$$\sum_{\tau=1}^{t} \hat{Y}_t \geq \sum_{\tau=1}^{t-L} (M - a_t \hat{P}_t - b_t L) \qquad \forall t \tag{3.25}$$

$$\sum_{\tau=1}^{t} \hat{Y}_t \leq \sum_{\tau=1}^{t-L} (M - a_t \hat{P}_t - b_t L) + \nu \qquad \forall t \tag{3.26}$$

$$\hat{X}'_t, \hat{P}_t, \hat{I}_t, \hat{Y}_t, \hat{R}_t \geq 0 \qquad \forall t \tag{3.27}$$

In the following section we examine the structure of locally optimal solutions to both the FLT and CF models to explore the differences between them, induced by the different models of production capacity they use.

Model Analysis

In Appendix 3.1, we show that the revenue function of the FLT model is concave, resulting in a concave objective function. Further, the linear demand function results in constraints (3.15)–(3.19) being linear. Thus the FLT model aims to maximize a concave function over a convex constraint set, so a locally optimal solution is also globally optimal. The CF model has a quasi-concave objective function if the sales variable is positive and the capacity constraint is tight (see Appendix 3.2). However, satisfying the capacity constraint at equality causes the constraint set to lose convexity and become concave. Hence the CF model does not have a unique global optimum. Nevertheless, all local optima should satisfy the Karush-Kuhn-Tucker (KKT) optimality conditions and since the global optimum must also be a local optimum, structural properties derived for a local optimum are valid for a global optimum.

We begin by examining the relationships between price, demand, lead time and capacity using the KKT conditions for a local optimum (Appendix 3.3). We then examine the relationship between the ending FGI and the delivery flexibility parameter ν and discuss properties of some Lagrange multipliers used in the formulations.

Sales, Price and Delivery Time at Optimality

We first develop expressions for price and sales quantity based on the KKT conditions for a local optimum. We are interested in local optima with non-trivial solutions, i.e., the firm operates in a reasonable manner that yields non-zero revenue, or in other words, both price and sales are non-zero. Using $P_t > 0$ in (3.46) and (3.61) we obtain the optimal prices for both models as follows:

Price(FLT model)

$$\hat{P}_t = \frac{M}{2a_t} - \frac{b_t}{2a_t}L - \frac{1}{2}\left(\hat{\mu}_t - \sum_{\tau=t+L}^{T} \hat{\sigma}_\tau + \sum_{\tau=t+L}^{T} \hat{\rho}_\tau\right) \tag{3.28}$$

Price(CF model)

$$P_t = \frac{M}{2a_t} - \frac{b_t}{2a_t}\left(\frac{\hat{W}_t}{X_t}\right) - \frac{1}{2}\left(\mu_t - \sum_{\tau=t+L_G}^{T} \sigma_\tau + \sum_{\tau=t+L_G}^{T} \rho_\tau\right) \tag{3.29}$$

Substituting these expressions into the demand functions for the respective models, we obtain the following expressions for the sales quantities:

Sales (FLT model)

$$\hat{D}_t = \frac{M}{2} - \frac{b_t}{2} - \frac{a_t}{2}\left(\sum_{\tau=t+L}^{T} \hat{\sigma}_t - \sum_{\tau=t+L}^{T} \hat{\rho}_t - \hat{\mu}_t\right) \tag{3.30}$$

Sales(CF model)

$$D_t = \frac{M}{2} - \frac{b_t}{2}\left(\frac{\hat{W}_t}{X_t}\right) - \frac{a_t}{2}\left(\sum_{\tau=t+L_G}^{T} \sigma_t - \sum_{\tau=t+L_G}^{T} \rho_t - \mu_t\right) \tag{3.31}$$

Equations (3.28)–(3.31) clearly show that under the CF model both price and sales decisions are dependent upon the observed lead time. Equation (3.29) is particularly interesting since price is expressed as a downward sloping function of lead time using the basic decision variables of the production system. Ray and Boyaci (2004) assume price to be a downward sloping function of lead time in order to investigate the effects of ignoring lead-time sensitivity of prices while making pricing decisions. However, our model does not require such an assumption, since the relationship between price and lead time emerges directly from the model. The last terms in all four expressions represent the interactions between the cumulative shipment constraints and can be interpreted in terms of the Lagrange multiplier of the finished goods inventory balance constraints of the respective models. We discuss this in the section "Properties of Lagrange Multipliers".

It is interesting to examine the behavior of the model as lead times approach the threshold delivery time L_G. We can write

$$\left(\frac{\hat{W}_t}{X_t}\right) = L_G - \left(L_G - \left(\frac{\hat{W}_t}{X_t}\right)\right) = L_G - \Delta L$$

which, in turn, allows us to rewrite (3.29) and (3.31) as:

$$P_t = \frac{M}{2a_t} - \frac{b_t}{2a_t}L_G + \frac{b_t}{2a_t}\Delta L - \frac{1}{2}\left(\mu_t - \sum_{\tau=t+L_G}^{T}\sigma_\tau + \sum_{\tau=t+L_G}^{T}\rho_\tau\right) \tag{3.32}$$

$$D_t = \frac{M}{2} - \frac{b_t}{2}L_G + \frac{b_t}{2}\Delta L - \frac{a_t}{2}\left(\sum_{\tau=t+L_G}^{T}\sigma_\tau - \sum_{\tau=t+L_G}^{T}\rho_\tau - \mu_t\right) \tag{3.33}$$

The ΔL term represents the difference between the maximum allowable delivery time and the average delivery time quotation (i.e., the average manufacturing lead time). When $\Delta L < 0$, i.e., the average delivery time quotation exceeds the maximum allowable delivery time, our model penalizes the firm by reducing demand per (3.33), thus reducing the WIP in the production system and hence the average lead time. This self-regulating behavior removes the need to include explicit penalty terms for exceeding the delivery time guarantee in the objective function of model CF.

This behavior can be visualized by examining relationships between different variables by means of a causal loop diagram (Sterman 2000) in Fig. 3.2. The variable at the tail of an arc is linked to the variable at the head of the arc by the sign on the head that indicates whether an increase in the variable at the tail causes a corresponding increase or a decrease in the variable at the head. Average delivery times eventually have a negative feedback on sales, since an increase in sales will cause an increase in quoted average delivery times, which in turn will reduce ΔL, making it negative. Negative values of ΔL cause a reduction in sales, keeping the average delivery time and sales variables in close relation with each other. Recall that these two variables are tightly coupled with the price variable through the demand function. Hence a reduction in both sales and average delivery time would require an increase in price.

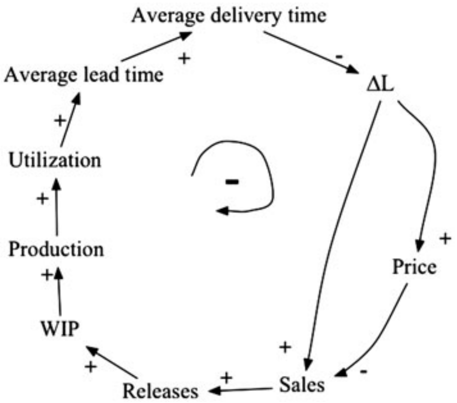

Fig. 3.2 Relationship between variables

Thus the firm can use both price and delivery time in an aggregate planning framework to manage sales. Webster (2002) proposes a similar feedback loop in his model for determining equilibrium values for price and lead time in face of a changing demand function. His model changes capacity in response to a change in sales, keeping lead time fixed, but does not consider the costs of changing capacity.

Prices and Utilization

If $\Delta L > 0$, the quoted average delivery time is less than the guaranteed delivery time. From (3.33), it would appear that this would cause sales to increase, increasing resource utilization and delivery times. Recall that a target guaranteed delivery time corresponds to a particular utilization level. However, it is not possible to determine a priori whether a targeted utilization level will allow satisfying the guaranteed delivery time, since we only have information about average delivery times and the maximum realized delivery time may exceed the guaranteed delivery time at high utilizations. Attaining higher utilization levels in the CF model requires additional WIP, and the marginal increase in utilization decreases with each unit increase in WIP. The CF model uses this information to determine a utilization level that may be lower than the target utilization level corresponding to the guaranteed delivery time parameter. This decision is implemented by limiting sales (to control utilization) by increasing prices instead ($\Delta L > 0$ in (3.32)). At high values of L_G, ΔL could be significant enough that a large difference may exist between the prices quoted by the two models.

Analytically, this behavior can be explained as follows. Let u_t denote the utilization level in period t due to production level X_t, i.e., $X_t = u_t K_1$. Using this relationship we obtain $L_t = \frac{\hat{W}_t}{K_1 u_t}$. Initially it appears counterintuitive that lead time decreases with increasing utilization. However, this expression must be viewed in the light of the relationship between WIP W_t and utilization u_t. As shown in Fig. 3.1, a unit increase in utilization will cause WIP, and thus lead times, to increase by a larger amount. Conversely, if utilization threatens to increase, the CF model can lower WIP, thus controlling it and keeping average delivery time in check. The FLT model is blind to the effects of utilization on delivery time and thus has one less lever for managing sales and system operation. The impact of this can be observed by simulating the key decisions of the FLT model in a congestion-prone system, which we do in the section "Low Utilization: $u = 0.8$, $L = 1$".

The key decisions made by a joint pricing and production planning model are prices and material releases. For the FLT model, these correspond to the variables $\hat{P}_t$ and $\hat{X}_t$ respectively. If we define $\hat{W}_t^{'}$ to be the WIP level arising from the material release $\hat{X}_t$, then the production in a given period is found to be $f(\hat{W}_t^{'})$ from the CF. Thus, the average delivery time can be written as $\frac{\hat{W}_t^{'}}{f(\hat{W}_t^{'})}$. Now let $D_t^{'}$ be the sales decision arising from this average delivery time quotation and price quotation $\hat{P}_t$. Then we have

$$D_t^{'} = M - a_t \hat{P}_t - b_t \left(\frac{\hat{W}_t^{'}}{f(\hat{W}_t^{'})} \right) \tag{3.34}$$

Further, if $D_t{>}0$ and $\hat{W}_t^{'} > 0$, then $\mu_t = 0$ and $\hat{\mu}_t = 0$ from complementary slackness conditions (3.52) and (3.65). Using this in (3.28) and (3.29), we can express the difference in prices as

$$\begin{aligned} P_t - \hat{P}_t &= \frac{b_t}{2a_t}\left(L - \frac{\hat{W}_t}{X_t}\right) + \sum_{\tau=t+L_G}^{T} (\sigma_t - \hat{\sigma}_t) - \sum_{\tau=t+L_G}^{T} (\rho_t - \hat{\rho}_t) \\ &= \frac{b_t}{2a_t}\Delta L^{'} + \sum_{\tau=t+L_G}^{T} (\sigma_t - \hat{\sigma}_t) - \sum_{\tau=t+L_G}^{T} (\rho_t - \hat{\rho}_t) \end{aligned}$$

At high values of L (corresponding to high utilization), it is possible that $\Delta \mathrm{L}^{'} > 0$ and is large enough for $\hat{P}_t$ to be significantly less than P_t. In this scenario, we have $D_t^{'} > D_t$. Further, since $\Delta L^{'} > 0$, the material release decisions $\hat{X}_t$ could load the system with significantly higher WIP than the release decisions R_t made by the CF model, leading to larger queue sizes. Average quoted lead times will not be met and there will not be enough FGI to satisfy sales $D_t^{'}$. If we allow unsatisfied sales to be lost, revenues will drop since the prices quoted are lower than those in the CF model. We discuss this further through a numerical example in the section "Experiments with Early Delivery Flexibility: $\nu > 0$".

On the other hand, at low values of L (corresponding to low utilization), $\Delta L^{'}$ will be small and will have less influence over the differences in prices quoted by the two models. From Fig. 3.1, it can be seen that lower utilizations imply lower WIP levels and hence lower average delivery times. Further, the marginal increase in throughput with a unit increase in WIP is higher at low utilization than at high utilization. This allows the production system to fulfill demand in a timely manner more easily. Lower utilizations are achieved by having low sales or ample excess capacity. Neither of these alternatives is practical in a capital-intensive environment, motivating our interest in high-utilization environments.

Properties of Lagrange Multipliers

Proposition 1 In an optimal solution to the CF model the capacity constraint is always tight (i.e., $\theta_t > 0$) if $\hat{W}_t, R_t, P_t$ and D_t are all strictly positive.

Proof See Appendix 3.4.

Capacity in the CF model is expressed in terms of the amount of WIP in the system that can be cleared in a given period. The above proposition implies that the release pattern will be coordinated with the sales pattern, so that there is just enough WIP to create the capacity required for fulfilling sales.

Proposition 2 In an optimal solution to the CF model, the marginal cost of holding finished goods inventory is always positive (i.e., $\pi_t > 0$) if $\theta_t > 0, X_t > 0$, and $\hat{W}_t > 0$.

Proof See Appendix 3.5.

Proposition 3 In an optimal solution to the FLT model, the marginal cost of holding finished goods inventory is always positive (i.e., $\gamma_t > 0$).

Proof See Appendix 3.6.

Since we quote an average delivery time, production that is realized earlier than due can be held as finished goods inventory to fulfill orders by the guaranteed delivery time. This can be clarified further when the marginal FGI costs are seen in relationship to the shipment Y_t. Considering the FLT model, if shipments $\hat{Y}_t > 0$, for some period t, then from condition (3.63), we have

$$\gamma_t - \sum_{\tau=t}^{T} \hat{\sigma}_\tau + \sum_{\tau=t}^{T} \hat{\rho}_\tau = 0$$

$$\Leftrightarrow -\sum_{\tau=t}^{T} \hat{\sigma}_\tau + \sum_{\tau=t}^{T} \hat{\rho}_\tau = -\gamma_t$$

If $Y_{t+L_G} > 0$, we have $-\sum\limits_{\tau=t}^{T} \hat{\sigma}_\tau + \sum\limits_{\tau=t}^{T} \hat{\rho}_\tau = -\gamma_{t+L_G}$

Since $\hat{P}_t > 0$, from condition (3.61), we have

$$-M + a_t\hat{P}_t + b_t L_G + a_t\left(\hat{P}_t + \hat{\mu}_t - \sum_{\tau=t+L_G}^{T} \hat{\sigma}_\tau + \sum_{\tau=t+L_G}^{T} \hat{\rho}_\tau\right) = 0$$

when $\hat{D}_t > 0, \hat{\mu}_t = 0$. Hence$-M + a_t\hat{P}_t + b_t L_G + a_t(\hat{P}_t - \gamma_{t+L_G}) = 0$. Rewriting, we obtain

$\gamma_{t+L_G} = 2\hat{P}_t - \frac{M}{a_t} + \frac{b_t}{a_t}L_G > 0$ by Proposition 3. Thus when there are positive sales in period t, it is beneficial to have FGI in period $t + L_G$ in order to meet the quoted delivery date. The analogous expression for the CF model is

$$\pi_{t+L_G} = 2P_t - \frac{M}{a_t} + \frac{b_t}{a_t}\left(\frac{\hat{W}_t}{X_t}\right).$$

Since $D_t,\ P_t$, and the average quoted delivery time for the CF model are strictly positive (section “CF model”), we find that the marginal cost of the FGI constraint in period $t + L$ is strictly positive by a simple manipulation of the demand function. In addition, the marginal cost of the FGI constraint in the CF model varies with both the price and the average delivery time, whereas for the FLT model it can vary only with price.

We also investigate the optimal sales decision made by our CF model if the linear demand function is replaced by the power function used by So and Song (1998). The demand function itself is expressed as $D_t = MP_t^{-a_t}L_t^{-b_t}$, where all symbols have the same meaning as before. By repeating the steps in Appendix 3.3 and the section “CF model”, we obtain the optimal sales decision as:

$$\ln D_t = \ln M - a_t \ln\left[\left(\frac{a_t}{1-a_t}\right)\left(\sum_{\tau=t+L_G}^{T}\rho_\tau - \sum_{\tau=t+L_G}^{T}\sigma_\tau\right)\right] - b_t \ln\left(\frac{\hat{W}_t}{X_t}\right) \quad (3.35)$$

We find that the negative feedback loop discussed in the section “Sales, Price, and Delivery Time at Optimality” for the linear demand function also holds for the power demand function, though on a logarithmic scale. We conjecture that the negative feedback relationship between sales and average delivery times would exist in case of any demand function form that is downward sloping in delivery time. We now present a numerical study to compare the behavior of the CF and FLT models.

Numerical Study

The length of the planning horizon is chosen to be 24 periods where each period corresponds to a month. The price and lead time sensitivities for each period are presented in Table 3.1. Price sensitivity is low in the first half of the horizon and increases in the latter half. This change in sensitivity represents a typical scenario in semiconductor products where, as other manufacturers bring competing devices to market, the price for the device will begin to decrease significantly (Akcali et al. 2000; Leachman and Ding 2007). Lead-time sensitivity, on the other hand, is low in the first and third quarters, and high in the second and fourth quarters of the horizon. The high-sensitivity periods represent seasonal effects where the market is unwilling to wait for a longer time interval between placing the order and taking delivery of the product.

Values of other input parameters are given in Table 3.2. The value of the curvature parameter K_2 is selected such that the slope of the CF at $\hat{W}_t = 0$ does not exceed the reciprocal of the raw process time.

Table 3.1 Price and lead-time sensitivities

Period range	Price sensitivity (a_t)	Lead-time sensitivity (b_t)
1–6	1	1
7–12	1	2
13–18	2	1
19–24	2	2

The relationship between utilization u and fixed lead time L for chosen values of the CF parameters K_1 and K_2 is obtained as seen in Appendix 3.7. We chose the value for guaranteed lead time as $L_G = L + 1$ periods. Thus sales will only be lost if the realized lead time exceeds the planned lead time L by more than one period. We consider four combinations of unit costs given in Table 3.3. Combination 1 is the base case. Combination 2 allows comparison of objective function values when unit material cost is less than the unit production cost. Combination 3 allows for a similar comparison when WIP holding cost is less than the FGI holding cost. To facilitate direct comparison of the objective function values, we use the modified FLT (MFLT) model that considered WIP costs in the objective function instead of the original FLT model used for the analytical results.

We assume there is no residual demand from earlier planning periods to be met in the current planning horizon. Both CF and modified FLT models are initialized with WIP equal to the targeted production in period 1, i.e., $W_0 = uK_1$. We also require that ending WIP in periods 23 and 24 for both CF and FLT model equals uK_1. WIP inventory in the FLT model at the end of a period is the sum of the releases in the previous L periods; we impose this boundary condition on the FLT model by controlling the material releases. By imposing these boundary conditions, we aim to avoid ramp-up and end-effects that would normally influence behavior at the beginning and end of the horizon.

Table 3.2 Input parameter values

Length of planning horizon	T	24 periods
Theoretical production capacity per period	K_1	500 units
Curvature parameter	K_2	100
Demand at zero price and zero lead time	M	1,000 units
Early delivery flexibility	ν	0 units
Fixed lead time L and corresponding target utilizations u	L	1 period ($u = 0.8$), 2 periods ($u = 0.9$), 4 periods ($u = 0.95$)
Initial WIP for CF and FLT models ($u = 0.8$)	W_o	400
WIP at ending of period 23 for CF and FLT models ($u = 0.8$)	W_{23}	400
WIP at ending of period 24 for CF and FLT models ($u = 0.8$)	W_{24}	400
Initial WIP for CF and FLT models ($u = 0.9$)	W_o	900
WIP at ending of period 23 for CF and FLT models ($u = 0.9$)	W_{23}	900
WIP at ending of period 24 for CF and FLT models ($u = 0.9$)	W_{24}	900
Initial WIP for CF and FLT models ($u = 0.95$)	W_o	1,900
WIP at ending of period 23 for CF and FLT models ($u = 0.95$)	W_{23}	1,900
WIP at ending of period 24 for CF and FLT models ($u = 0.95$)	W_{24}	1,900

Table 3.3 Unit cost combinations

Combination	Unit material cost	Unit production cost	Unit WIP holding cost	Unit FGI holding cost
1	1/unit	1/unit	1/unit/period	1/unit/period
2	0.5/unit	1/unit	1/unit/period	1/unit/period
3	1/unit	1/unit	0.5/unit/period	1/unit/period
4	0.5/unit	1/unit	0.125/unit/period	0.25/unit/period

Both models are solved using the CONOPT solver in the general algebraic modeling system (GAMS) optimization suite (www.gams.com). Since this solver does not guarantee a globally optimal solution for the nonconvex CF model, we used six different starting points for both models, and found that for both models all initial starting points led to the same values for the objective function and decision variables.

The primary question of interest is how important it is to consider the effects of congestion explicitly—do they lead to significant differences in profit, and, if so under what conditions? One way to approach this issue is to examine how much profit-planned solutions from the FLT model would yield if the production system is subject to the type of congestion represented in the CF model. In other words, how much profit is lost if we plan using a fixed lead time when our production system is, in reality, subject to congestion as represented by a CF?

In order to examine this question, we simulate the behavior of the production system period by period using expressions (3.6) and (3.7). The material releases obtained from the MFLT model are used to determine the WIP level in each period, and the CF is used to determine the production at this WIP level. The WIP level in a period can be calculated using the ending WIP in the previous period and the release in the current period using expressions derived in Appendix 3.8. We estimate the average delivery time resulting from the WIP level as $L_t = \hat{W}_t / X_t$. These provide estimates of the realized production, WIP and finished goods inventory available when the system operates as represented by the CF, allowing us to obtain actual shipments. Projected sales in each period under the price quoted by the FLT model are given by the linear demand function, allowing us to calculate the revenue that would be realized if the production system were able to produce exactly the quantities planned by the FLT model in each period. We assume sales are lost if not enough finished goods are available for order fulfillment. The realized shipments are multiplied by the quoted price to give the realized revenue in each period. We deduct the material, production and inventory costs incurred due to the release and sales decisions to obtain the realized profit for both models. We impose no boundary conditions on the system during this simulation.

Experiments without Early Delivery Flexibility: $\nu = 0$ In our base case we use the unit costs described by Combination 1 to study the behavior of each model with no early delivery flexibility, i.e., $\nu = 0$. We will discuss the results for each planned utilization level u, and hence each planned lead time L, separately. In all figures, the captions "FLT," and "CF" denote the quantity computed by the respective optimization models. "Realized FLT" denotes the quantities that are realized when the plans computed from the FLT model are implemented in a system that is subject to congestion as represented by the CF.

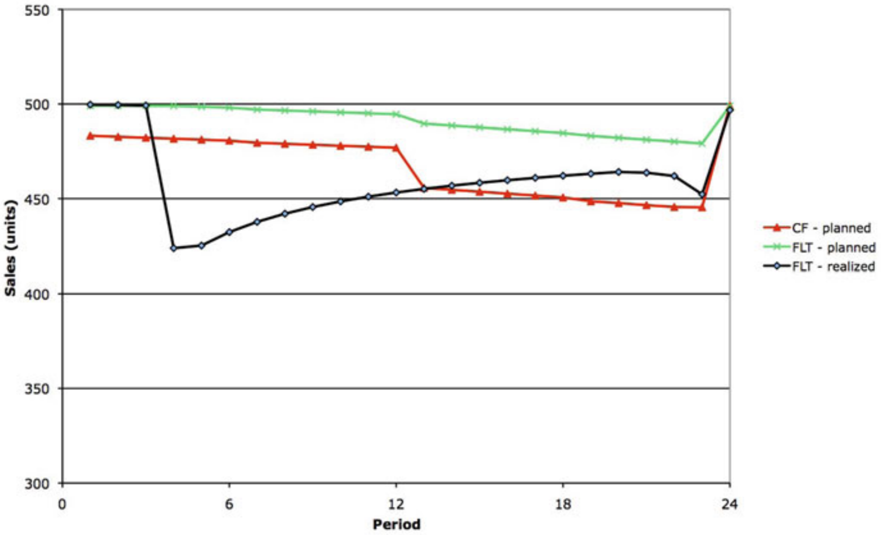

Fig. 3.3 Sales comparison at $u = 0.8$

Low utilization: $u = 0.8$*,* $L = 1$ The results of this experiment are summarized in Figs. 3.3, 3.4 and 3.5. As seen in Fig. 3.4, the CF model consistently sets prices somewhat higher than MFLT model, but not by a great deal. Both models reduce prices in the second half of the planning horizon when the market becomes sensitive to price. However, Fig. 3.3 shows that the FLT model realizes substantially lower sales than the CF model in the later periods. Examination of Figs. 3.5 and 3.6, which show the planned lead times and FGI levels, explains the situation. The CF model plans to operate at a higher utilization level with longer lead times from the start of the horizon. It must meet demands within the maximum lead time L_G, but accomplishes this by building finished goods inventory early in the horizon which it draws down over time, allowing the model to meet demand within the specified maximum lead time L_G that the market will bear. As a result of this approach and the slightly higher prices it sets, the planned sales of the CF model are lower than those of the FLT model.

However, Fig. 3.6 shows that the finished goods inventory realized when the material releases and prices from the FLT plan are implemented in the presence of congestion is very different from that planned. The FLT model assumes that any demand that does not exceed the capacity of the system can be met within the planned lead time $L = 1$, allowing it to set lower prices than CF. However, the low prices and low-quoted lead time lead to high demand, which the congested system cannot meet within the planned lead time L_G. This results in a stock out in periods 9 through 11 where there is no available product to ship and sales are lost. The net result, seen in Fig. 3.7, is an approximately 20 % difference in planned and realized revenue for the FLT model in periods 11 through 20.

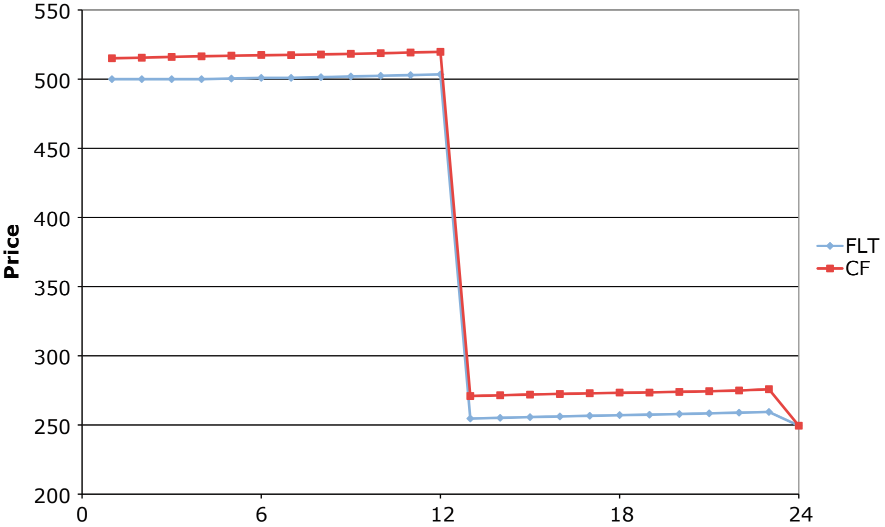

Fig. 3.4 Pricing comparison at $u = 0.8$

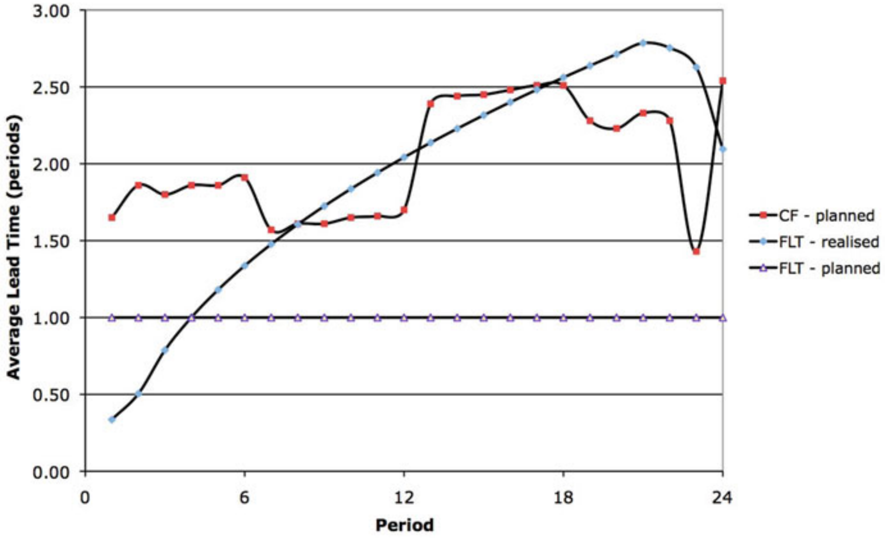

Fig. 3.5 Planned and realized average lead times at $u = 0.8$

Figure 3.7 shows that both models plan to achieve very similar revenue, but the CF is able to achieve its aim while the FLT model is not. The difference is almost entirely due to the FLT model's assumption that the fixed lead time L can be maintained regardless of utilization. The FLT model loads the system to its available theoretical capacity, which results in utilization levels incompatible with the maximum lead time

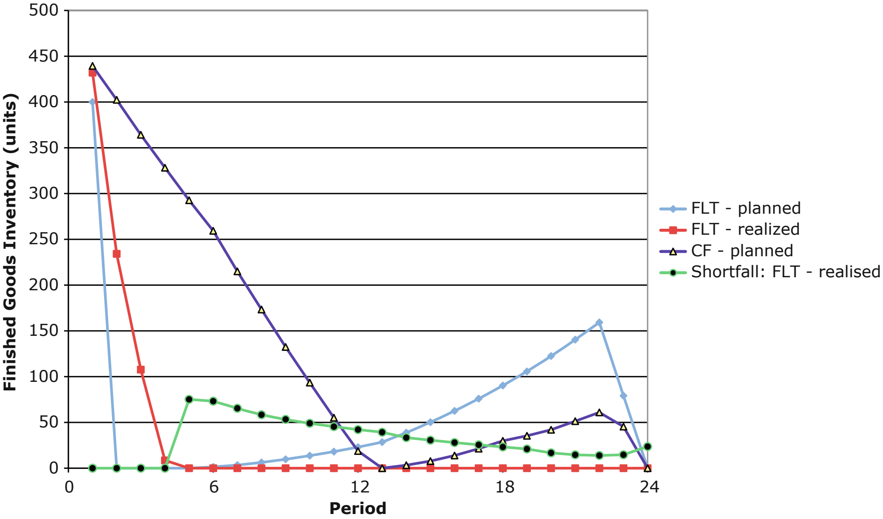

Fig. 3.6 Finished goods inventory levels for $u = 0.8$

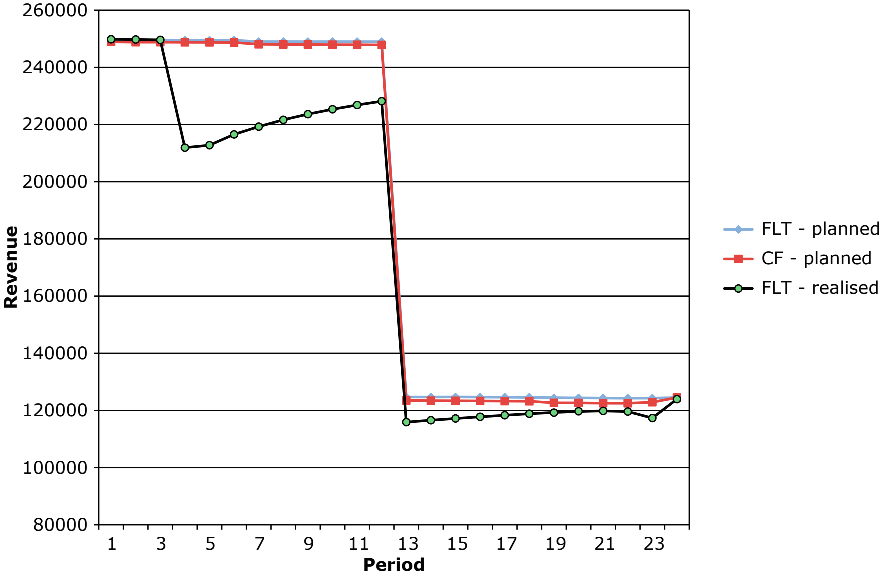

Fig. 3.7 Revenue comparison for u = 0.8

L_G. It is also interesting that this significant difference in behavior occurs despite low demand sensitivity to both prices and lead times.

The results of this experiment highlight what we believe is the principal reason for an FLT model to perform poorly in an environment subject to congestion. The basic

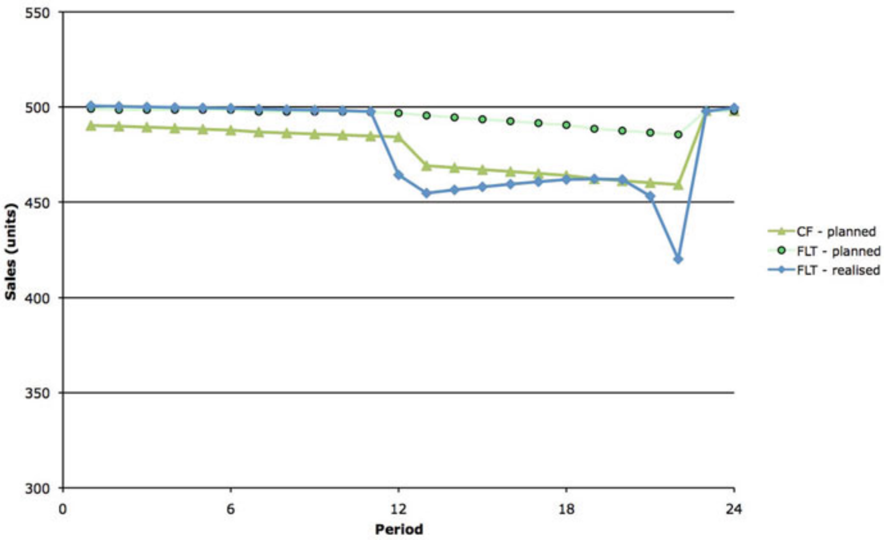

Fig. 3.8 Sales comparison for $u = 0.9$

issue is that the planning model fails to represent accurately the realized behavior of the production system, which is manifested in the realized lead time. Figure 3.5 clearly indicates that the planned lead time L is a gross underestimate of the realized lead time that becomes worse as the planning horizon advances.

Intermediate utilization: $u = 0.9,\ L = 2,\ L_G = 3$ The results of this experiment are given in Figs. 3.8–3.11. In this situation the difference between the two models is rather less than might be expected, although the behavior of the inventory and lead times differs somewhat between the models. This is because the maximum lead time L_G is consistent with a high level of utilization. The FLT model again loads the system to its capacity, resulting in lead times higher than L_G, but because L_G is already quite high the impact on predicted lead times is not as severe as at the lower utilization level. The CF model, on the other hand, varies lead times over the horizon, keeping them below L_G. Hence in this case both models plan very similar total revenues and both achieve them, although with quite different production plans. The reason both are able to achieve their plans to a large extent is the low sensitivity of demand to prices and lead times.

High utilization level: $u = 0.95,\ L = 4,\ L_G = 5$ In this case, again the difference between the two models is closer than before (Figs. 3.12–3.14). The primary reason for this is the high WIP level imposed at the beginning of the horizon for both models. Both models behave similarly, choosing not to make any releases into the system in the first few periods and consuming the initial WIP. This allows lead times to be low for both the CF and the realized FLT decisions in this initial part of the horizon. The FLT model again loads the system to capacity in the following periods, due to which the realized lead times gradually rise over the horizon. This is the only case

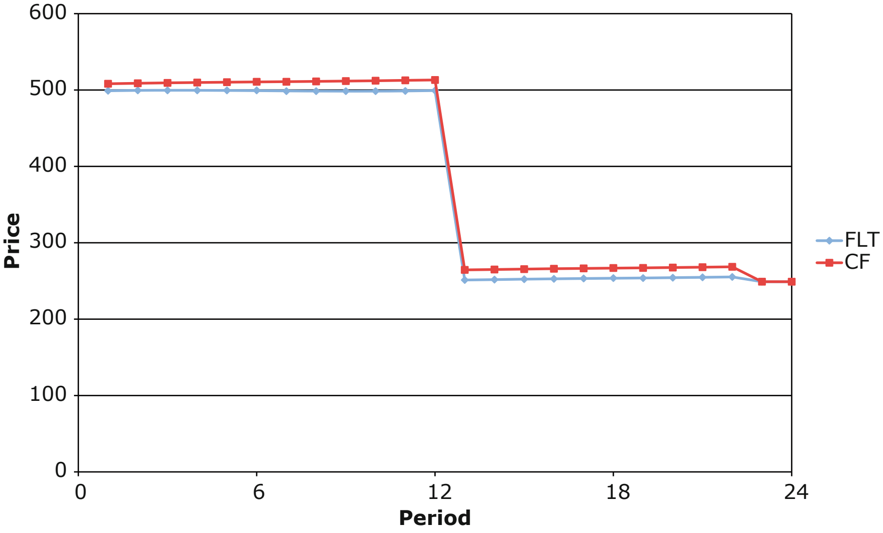

Fig. 3.9 Price comparison for $u = 0.9$

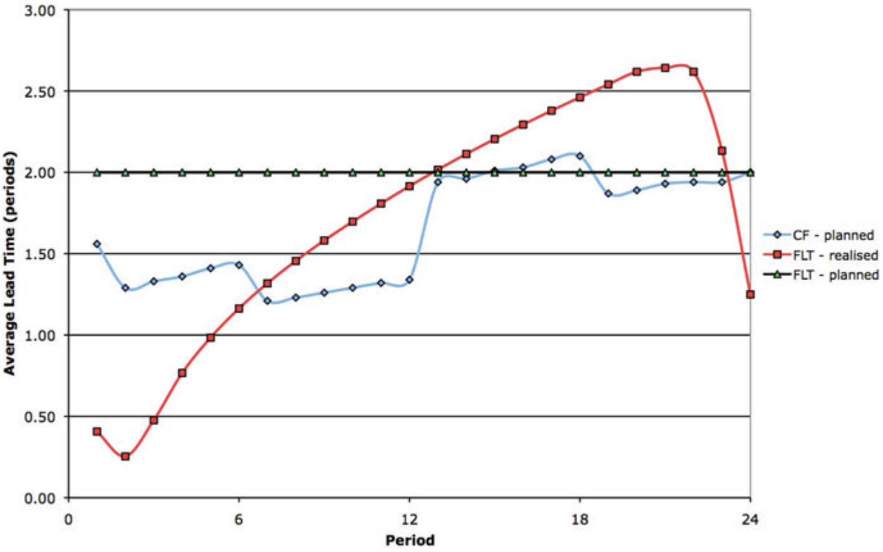

Fig. 3.10 Lead time comparison for $u = 0.9$

where we impose the ending WIP conditions on the simulated decisions of the FLT model, because otherwise the ending WIP does not rise to a value high enough to satisfy the ending conditions. This is again due to the fact that there are no releases early in the horizon, resulting in low WIP levels that do not rise fast enough during

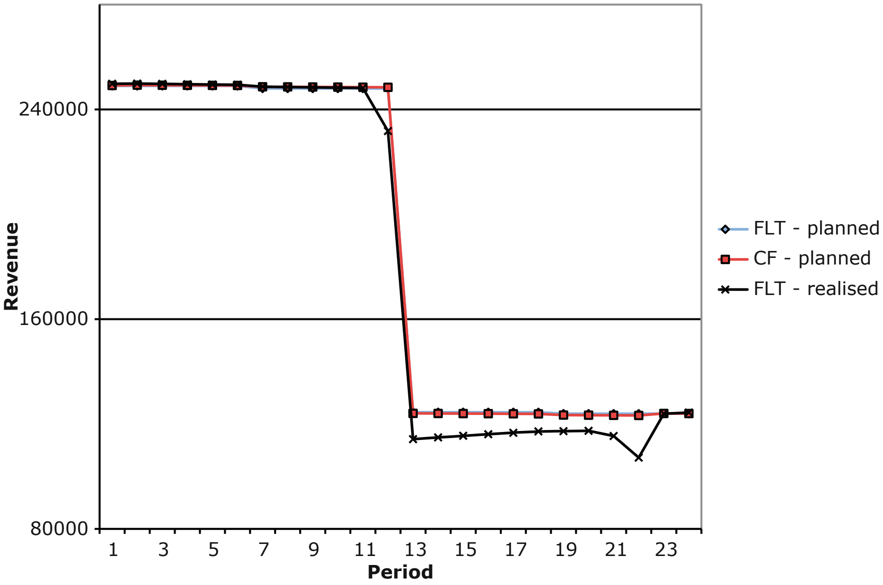

Fig. 3.11 Revenue comparison for $u = 0.9$

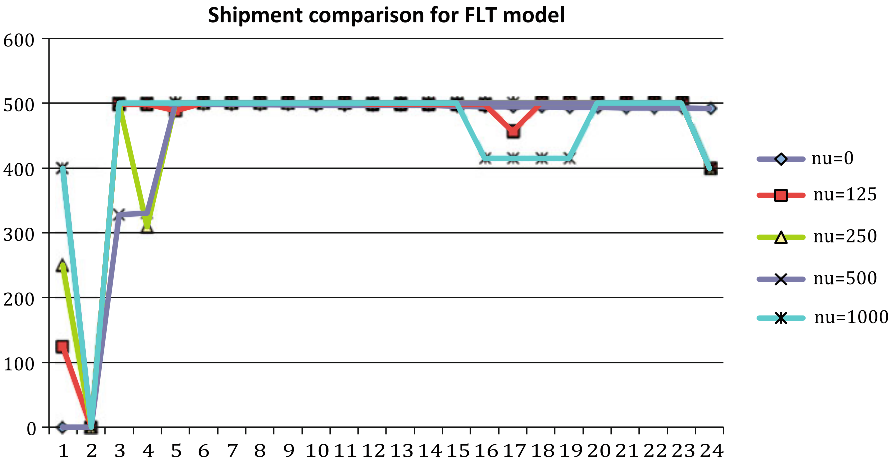

Fig. 3.12 Shipment decision comparisons for FLT model with changing ν

the course of the horizon. The detailed results of this experiment are omitted for the sake of brevity.

Objective function values The discussion to this point has demonstrated that the production and pricing plans developed by the CF and FLT models result in quite different plans over the planning horizon. When the planned lead time substantially

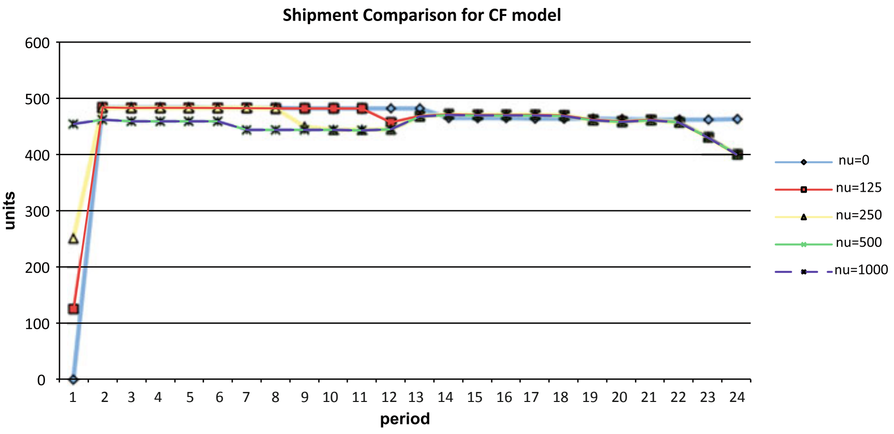

Fig. 3.13 Shipment decision comparisons for CF model with changing ν

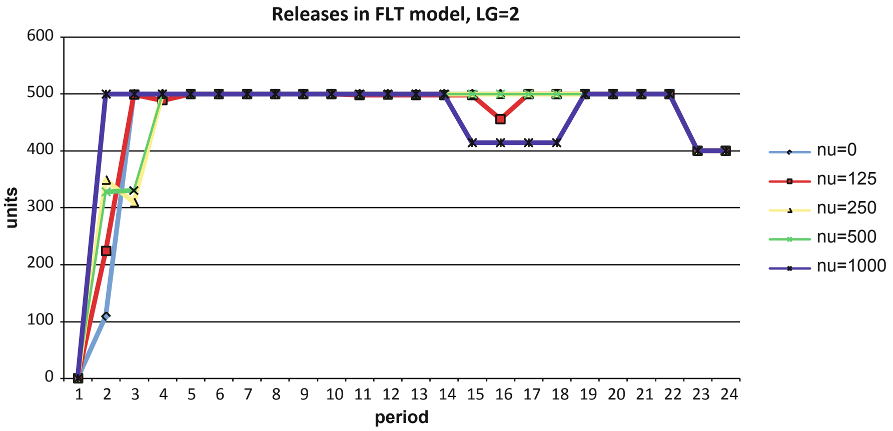

Fig. 3.14 Release decision comparisons for FLT model with changing ν

underestimates the manufacturing lead time that can actually be realized, severe discrepancies between the plans and actual deliveries to customers can result, as was the case for our experiment with $u = 0.8$. Table 3.4 presents a comparison of the realized objective function values planned by the FLT model, and those realized when the FLT plans are simulated in the presence of congestion. All quantities are expressed as a ratio to the objective function value obtained by the CF model for the same cost combination.

When $u = 0.8$, we find that even though the FLT model predicts an objective function value higher than the CF model, the realized objective (Fixed Lead Time-Simulated FLT-SIM) is significantly lower than both FLT and CF models. This is

Table 3.4 Comparison of planned and realized objective function values

(u, L, L_G)	Cost scenario (c_t, f_t, h_t, w_t)	FLT	FLT-SIM
(0.8, 1,2)	(1,1,1,1)	1.065	0.930
	(0.5, 1,1,1)	1.065	0.930
	(1,1,0.5,1)	1.064	0.930
	(0.5,1,0.125,0.25)	1.062	0.930
(0.9, 2,3)	(1,1,1,1)	0.999	0.978
	(0.5, 1,1,1)	0.999	0.978
	(1,1,0.5,1)	1.000	0.978
	(0.5,1,0.125,0.25)	0.999	0.977
(0.95, 4, 5)	(1,1,1,1)	0.987	0.993
	(0.5, 1,1,1)	0.988	0.993
	(1,1,0.5,1)	0.990	0.993
	(0.5,1,0.125,0.25)	0.991	0.994

due to the release decisions proposed by the FLT model that result in high WIP levels, high lead times and product shortages, all of which lead to lower revenue and profit margin. When the discrepancy between planned and realized lead times is less severe, when $u = 0.9$, the same effect is observed although at a much smaller level. When $u = 0.95$, the FLT-SIM is very slightly higher than the planned FLT objective, because the realized lead time is shorter than the planned lead time for most of the planning horizon. It is notable that the CF model gives the highest objective function value in all scenarios considered, most markedly when the discrepancy between planned and realized FLT lead time is most severe.

Experiments with Early Delivery Flexibility: $\nu > 0$ The combinations with early delivery flexibility ν provide more interesting insights. Early delivery flexibility allows both plans to shift production away from periods with high-delivery-time sensitivity to those with low-delivery-time sensitivity without the need to carry all the production as finished goods inventory. This should result in an increase in profit margins as flexibility increases, due to reduction in cost of carrying finished goods inventory. The two models use this flexibility differently. The shipment decisions made by the FLT model for different values of ν and cost Combination 4 when $L_G = 2$ are seen in Fig. 3.12 and those for the CF model in Fig. 3.13. The FLT model applies all of its flexibility in the early part of the horizon, choosing to make zero shipments in period 2. The model also chooses to increase the load in the system as ν increases by releasing more orders, which has a detrimental effect on the profit margin when its decisions are subjected to congestion.

Table 3.5 summarizes the planned and realized objective function values of the models, again using the objective function value of the CF model as a base. We observe that the profit margin for FLT-SIM decreases as ν increases from 125 to 500. The CF model also uses its flexibility early on for lower values of ν, but for $\nu = 500$ and 1,000, it spreads this flexibility over the horizon. It is interesting to note that the realized objective function value FLT-SIM first decreases and then increases with ν, suggesting that the choice of an optimal value for ν may improve the realized performance of the FLT model. However, it is again striking that the CF model produces higher objective function values consistently across all scenarios.

Table 3.5 Objective function comparison for experiments with early delivery

(u, L, L_G)	n	FLT	FLT-SIM
(0.8, 1,2)	0	1.062	0.930
	125	1.062	0.938
	250	1.062	0.935
	500	1.062	0.934
	1,000	1.062	0.962
(0.9, 2,3)	0	0.999	0.977
	125	0.999	0.984
	250	0.999	0.977
	500	0.998	0.993
	1,000	0.998	0.997
(0.95, 4, 5)	0	0.991	0.994
	125	0.991	0.996
	250	0.990	0.995
	500	0.990	0.995
	1,000	0.990	0.995

Conclusions and Future Directions

In this chapter we have used the concept of CFs developed in the production planning literature to develop an integrated model for jointly planning production and pricing over time for a manufacturing firm whose resources are subject to congestion. Our analytical results show that the interplay between lead times and prices in the demand function requires careful consideration of the implications of pricing decisions for lead times. Pricing decisions made under a naïve capacity model that assumes any level of demand up to the theoretical capacity of the system can be met within a fixed lead time independent of workload have the potential to lead to significant difficulties when low prices and optimistic lead time estimates lead to the system being unable to meet demand within a reasonable time, causing lost sales and possibly loss of customer goodwill. It is interesting that noticeable effects can be observed even when the demand is not very sensitive to prices or lead times.

The critical issue is the difference between the lead times assumed in the planning model and the realized lead times. A FLT model may perform satisfactorily in terms of achieving its planned revenue even at high utilization if the planned lead time is set consistently with the realized utilization levels and remains within the maximum lead time the market will bear. However, such a model will have difficulties when lead times are underestimated or when sensitivity to lead times changes abruptly, since it has no ability to modulate the lead times quoted based on system state and market sensitivities. It is also noteworthy that the CF model consistently sets higher prices than the FLT model, which upon reflection is intuitive; the price set by the CF model considers the costs incurred due to congestion such as WIP accumulation, whereas the FLT model does not. When solved at an aggregate level considering product families and planning horizons of 18–24 months, the models can be solved sufficiently rapidly to permit extensive what-if analysis to provide decision makers with intuition as to the likely results of their decisions.

The CF model is, as far as we are aware, the first model to integrate dynamic pricing and production planning over time in a manner that represents the effects of congestion due to queuing. Most queuing-based models provide steady-state results, while most prior models that plan prices and production over time have adopted conventional models of capacity that do not capture the effect of workload on lead times, and do not permit the joint manipulation of lead times and prices to maximize profit.

These results highlight the importance of a well-designed and functional manufacturing-marketing interface for firms operating in markets where price and lead-time sensitivity may change over time. The problem is aggravated by the fact that lead times are generally the responsibility of the supply chain organization, while pricing is determined by sales and marketing groups. A common solution to this issue we have observed in industry, and which has been advocated by a number of authors (Graves 1986; de Kok and Fransoo 2003) is to simplify the situation by requiring the supply chain organization to maintain a constant lead time which is agreeable to marketing. However, in capital intensive industries where resources must be run at high utilization for the firm to be profitable, small changes in utilization make maintaining a constant lead time a very challenging task. The CF model proposed here in fact addresses exactly this—modulating prices and releases to optimize profit within the constraints of the lead time imposed by the market's "reservation" lead time L_G. In addition, the ability to change both prices and lead times in response to changing market sensitivity may result in higher revenues and profits relative to using price as the only control variable.

Several directions for future work present themselves. A natural direction is the extension of the models developed in this chapter to environments with multiple product families that may serve quite different markets but share capacity. Many semiconductor wafer fabs operating as foundries produce circuits for quite different markets, such as controllers and communication devices, in the same fab using largely the same technology and equipment. Another natural extension is to embed these models in a multistage stochastic programming framework where scenarios would consider different price sensitivities for different products over time. This model presents a number of challenges due to the rapid growth of the scenario tree, but may still be practical for the aggregate models of the type suggested in this chapter, and considered by Allison et al. (1997).

Appendix 3.1: Concavity of Revenue Function for the FLT Model

Our revenue function has the form $R = PD = P(M - aP - bL) = MP - aP^2 - bLP$. Thus there is only one variable, P. Taking the second derivative of the revenue function w.r.t P, we obtain

$$\frac{d^2R}{dP^2} = -2a \leq 0$$

Since we assume $a \geq 1$,we have $-2a < 0$. Hence the revenue function is strictly concave.

Appendix 3.2: Nature of Demand and Revenue Function for the CF Model

The demand function has the form. $D_t = M - a_t P_t - b_t \left(\frac{\hat{W}_t}{X_t}\right)$. Under very reasonable conditions (see Proposition 3), we can show that the capacity constraint is tight, i.e., $X_t = f(\hat{W}_t) = \frac{K_1 \hat{W}_t}{K_2 + \hat{W}_t}$. Under these conditions, the demand function takes the form $D_t = M - a_t P_t - \frac{b_t}{K_1}(K_2 + \hat{W}_t)$. Thus, the demand function has a linear form. The CF constraint (3.8) is convex, hence the constraint set is also convex.

Then, revenue function has the form $g_t = P_t D_t = P_t (M - a_t P_t - \frac{b_t}{K_1}(K_2 + \hat{W}_t))$. We have:

$$\frac{\partial g_t}{\partial \hat{W}_t} = -\frac{b_t}{K_1} P_t; \frac{\partial^2 g_t}{\partial \hat{W}_t^2} = 0$$

$$\frac{\partial g_t}{\partial P_t} = M - 2a_t P_t - \frac{b_t}{K_1}(K_2 + \hat{W}_t); \frac{\partial^2 g_t}{\partial {P_t}^2} = -2a_t$$

$$\frac{\partial^2 g_t}{\partial P_t \partial \hat{W}_t} = -\frac{b_t}{K_1}$$

In order to have quasi-concavity, we require that

$$-\left(\frac{\partial g_t}{\partial \hat{W}_t}\right)^2 = -\left(-\frac{b_t}{K_1} P_t\right)^2 = -\frac{{b_t}^2}{K_1^2} P_t^2 \tag{3.36}$$

and

$$2 \frac{\partial^2 g_t}{\partial \hat{W}_t \partial P_t} . \frac{\partial g_t}{\partial \hat{W}_t} . \frac{\partial g_t}{\partial P_t} - \frac{\partial^2 g_t}{\partial \hat{W}_t^2} \left(\frac{\partial g_t}{\partial \hat{W}_t}\right)^2 - \frac{\partial^2 g_t}{\partial {P^2}_t} \left(\frac{\partial g_t}{\partial \hat{W}_t}\right) > 0$$

After some algebra the expression above reduces to $2 \frac{b_t^2}{K_1^2} P_t \left(M - a_t P_t - \frac{b_t}{K_1}(K_2 + \hat{W}_t)\right)$. If sales $D_t = M - a_t P_t - \frac{b_t}{K_1}(K_2 + \hat{W}_t))\rangle 0$, it is clear that this expression is nonnegative. Thus from (3.36) and this expression we conclude that the revenue function is quasi-concave.

Appendix 3.3: KKT Conditions for CF and FLT Models

KKT conditions for CF model The Lagrangian for this planning model is as below:

$$L = -\sum_{t=1}^{T} \left[MP_t - a_t P^2 - b_t P_t \left(\frac{\hat{W}_t}{X_t}\right) - c_t R_t - \phi_t X_t - h_t I_t - \omega_t W_t \right]$$

$$
\begin{aligned}
&+\sum_{t=1}^{T}\lambda_t(W_t - W_{t-1} + X_t - R_t) + \sum_{t=1}^{T}\pi_t(I_t - I_{t-1} - X_t + Y_t)\\
&+\sum_{t=1}^{T}\theta_t(K_2X_t + X_t\hat{W}_t) + \sum_{t=1}^{T}\mu_t\left(-M + a_tP_t + b_t\left(\frac{\hat{W}_t}{X_t}\right)\right)\\
&+\sum_{t=1}^{T}\pi_t\left(-\sum_{\tau=1}^{t}Y_\tau + \sum_{\tau=1}^{\tau=t-L_G}\left(-M + a_\tau P_\tau + b_\tau\left(\frac{\hat{W}_\tau}{X_\tau}\right)\right)\right)\\
&+\sum_{t=1}^{T}\rho_t\left(\sum_{\tau=1}^{t}Y_\tau - \sum_{\tau=1}^{\tau=t-L_G}\left(-M + a_\tau P_\tau + b_\tau\left(\frac{\hat{W}_\tau}{X_\tau}\right)\right) - \upsilon\right)\\
&+\sum_{t=1}^{T}\chi_t\left(\hat{W}_t - \frac{1}{2}(W_{t-1} + W_t)\right)
\end{aligned}
$$

The first order optimality conditions and complementary slackness conditions follow:

First Order Optimality Conditions

$$
\begin{aligned}
\frac{\partial L}{\partial I_t} &= h_t + \pi_t - \pi_{t-1} \geq 0\\
\frac{\partial L}{\partial I_t} &= h_T + \pi_T
\end{aligned}
\tag{3.37}
$$

$$
\begin{aligned}
I_t\frac{\partial L}{\partial I_t} &= I_T(h_T + \pi_t - \pi_{t-1}) = 0\\
I_t\frac{\partial L}{\partial I_t} &= I_T(h_T + \pi_T)
\end{aligned}
\tag{3.38}
$$

$$
\frac{\partial L}{\partial R_t} = c_T - \lambda_t \geq 0 \tag{3.39}
$$

$$
R_t\frac{\partial L}{\partial R_t} = R_T(c_T - \lambda_T) = 0 \tag{3.40}
$$

$$
\frac{\partial L}{\partial W_t} = \omega_t + \lambda_t - \lambda_{t+1} - \frac{1}{2}(\chi_t + \chi_{t+1}) \geq 0 \tag{3.41}
$$

$$
W_t\frac{\partial L}{\partial W_t} = 0 \tag{3.42}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial X_t} &= \phi_t + \lambda_t - \pi_t + \theta_t(K_2 + \hat{W}_t)\\
&\quad - \frac{b_t\hat{W}_t}{X_t^2}\left(P_t + \mu_t - \sum_{\tau=t+L_G}^{T}\sigma_\tau + \sum_{\tau=t+L_G}^{T}\rho_\tau\right)
\end{aligned}
\tag{3.43}
$$

$$X_t \frac{\partial L}{\partial X_t} = 0 \tag{3.44}$$

$$\frac{\partial L}{\partial P_t} = -M + a_t P_t + b_t \left(\frac{\hat{W}_t}{X_t} \right) + a_t \left(P_t + \mu_t - \sum_{\tau = t + L_G}^{T} \sigma_\tau + \sum_{\tau = t + L_G}^{T} \rho_\tau \right) \geq 0 \tag{3.45}$$

$$P_t \frac{\partial L}{\partial P_t} = 0 \tag{3.46}$$

$$\frac{\partial L}{\partial Y_t} = \pi_t - \sum_{\tau = t}^{T} \sigma_\tau + \sum_{\tau = t}^{T} \rho_\tau \geq 0 \tag{3.47}$$

$$Y_t \frac{\partial L}{\partial Y_t} = 0 \tag{3.48}$$

$$\frac{\partial L}{\partial \hat{W}_t} = \theta_t (X_t - K_1) + \chi_t + \frac{b_t}{X_t} \left(P_t + \mu_t - \sum_{\tau = t + L_G}^{T} \sigma_\tau + \sum_{\tau = t + L_G}^{T} \rho_\tau \right) \geq 0 \tag{3.49}$$

$$\hat{W}_t \frac{\partial L}{\partial \hat{W}_t} = 0 \tag{3.50}$$

Complementary Slackness Conditions

$$\theta_t (K_2 X_t + X_t \hat{W}_t - K_t \hat{W}_t) \tag{3.51}$$

$$\mu_t \left(-M + a_t P_t + b_t \left(\frac{\hat{W}_t}{X_t} \right) \right) = 0 \tag{3.52}$$

$$\sigma_t \left(-\sum_{\tau = 1}^{t} Y_\tau + \sum_{\tau = 1}^{t - L_G} \left(M - a_\tau P_\tau - b_\tau \left(\frac{\hat{W}_\tau}{X_\tau} \right) \right) \right) = 0 \tag{3.53}$$

$$\rho_t \left(\sum_{\tau = 1}^{t} Y_\tau - \sum_{\tau = 1}^{t - LG} \left(M - a_\tau P_\tau - b_\tau \left(\frac{\hat{W}_\tau}{X_\tau} \right) \right) - v \right) = 0 \tag{3.54}$$

$$\chi_t \left(\hat{W}_t - \frac{1}{2} (W_{t-1} + W_t) \right) = 0 \tag{3.55}$$

Nonnegativity conditions λ_t, π_t: unrestricted, $\theta_t, \mu_t, \sigma_t, \rho_t, \chi_t \geq 0$

KKT conditions for FLT model The Lagrangian for this planning model is as below:

$$\begin{aligned} L = & -\sum_{\tau = 1}^{t} \left[M \hat{P}_t - a_t \hat{P}_t^2 - b_t \hat{P}_t L_G - c_t \hat{X}_t - h_t \hat{I}_t \right] \\ & + \sum_{\tau = 1}^{t} \gamma_t \left(\hat{I}_t - \hat{I}_{t-1} - \hat{X}_{t - L_G} + \hat{Y}_t \right) \end{aligned}$$

$$+\sum_{t=1}^{T}\delta_t(\hat{X}_t - K_1) + \sum_{t=1}^{T}\hat{\mu}_t(-M + a_t\hat{P}_t + b_tL_G)$$

$$+\sum_{t=1}^{T}\hat{\sigma}_t\left(-\sum_{\tau=1}^{t}\hat{Y}_t + \sum_{T=1}^{t-L_G}(M - a_\tau\hat{P}_\tau - b_\tau L_G)\right)$$

$$+\sum_{t=1}^{T}\hat{\rho}_t\left(\sum_{\tau=1}^{t}\hat{Y}_\tau - \sum_{T=1}^{t-LG}(M - a_\tau\hat{P}_\tau - b_\tau L_G) - v\right)$$

The first order optimality conditions and complementary slackness conditions follow.

First Order Optimality Conditions

$$\frac{\partial L}{\partial \hat{I}_t} = h_t + \gamma_t - \gamma_{t+1} \geq 0 \tag{3.56}$$

$$\hat{I}t\frac{\partial L}{\partial \hat{I}_t} = 0 \tag{3.57}$$

$$\frac{\partial L}{\partial \hat{X}_t} = c_t - \gamma_{t+L_G} + \delta_t \geq 0 \tag{3.58}$$

$$\hat{X}_t\frac{\partial L}{\partial \hat{X}_t} = 0 \tag{3.59}$$

$$\frac{\partial L}{\partial \hat{P}_t} = -M + a_t\hat{P}_t + b_tL + a_t\left(\hat{P}_t + \hat{\mu}_t - \sum_{\tau=t+L}^{T}\hat{\sigma}_\tau + \sum_{\tau=t+L}^{T}\hat{\sigma}_\tau\right) \geq 0 \tag{3.60}$$

$$\hat{P}_t\frac{\partial L}{\partial \hat{P}_t} = 0 \tag{3.61}$$

$$\frac{\partial L}{\partial Y_t} = \gamma_t - \sum_{l=t}^{T}\hat{\sigma}_l + \sum_{l=t}^{T}\hat{P}_l \geq 0 \tag{3.62}$$

$$\hat{Y}_t\frac{\partial L}{\partial \hat{Y}_t} = 0 \tag{3.63}$$

Complementary Slackness Conditions

$$\delta_t(\hat{X}_t - K) = 0 \tag{3.64}$$

$$\hat{\mu}_t(-M + a_t\hat{P}_t + b_tL_G) = 0 \tag{3.65}$$

$$\hat{P}(\sum_{l=1}^{t}\hat{Y}_l + \sum_{l=1}^{t-L_G}(M - a_l\hat{P}_l - b_lL_G)) = 0 \tag{3.66}$$

$$\hat{P}_t(\sum_{l=1}^{t}\hat{Y}_l - \sum_{l=1}^{t-L_G}(M - a_l\hat{P} - b_lL_G) - v) = 0 \tag{3.67}$$

Nonnegativity conditions γ_t: unrestricted, δ_t, $\hat{\mu}_t$, $\hat{\sigma}_t$, $\hat{\rho}_t \geq 0$

Appendix 3.4: Proof of Proposition 1

From equation (3.50), we have $\hat{W}_t \frac{\partial L}{\partial \hat{W}_t} = 0$.

From $\hat{W}_t > 0$, we have

$$\frac{\partial L}{\partial \hat{W}_t} = \theta_t(X_t - K_1) + \chi_t + \frac{b_t}{X_t}\left(P_t + \mu_t - \sum_{\tau=t+L_G}^{T} \sigma_\tau + \sum_{\tau=t+L_G}^{T} \rho_\tau\right) = 0$$

$$\Rightarrow \theta_t = \frac{1}{(K_1 - X_t)}\left(\chi_t + \frac{b_t}{X_t}\left(P_t + \mu_t - \sum_{\tau=t+L_G}^{T} \sigma_\tau + \sum_{\tau=t+L_G}^{T} \rho_\tau\right)\right) \tag{3.68}$$

From equation (3.46), we have $P_t \frac{\partial L}{\partial P_t} = 0$.

From $P_t > 0$,

$$\frac{\partial L}{\partial P_t} = -M + a_t P_t + b_t\left(\frac{\hat{W}_t}{X_t}\right) + a_t\left(P_t + \mu_t - \sum_{\tau=t+L_G}^{T} \sigma_\tau + \sum_{\tau=t+L_G}^{T} \rho_\tau\right) = 0$$

$$\Rightarrow \left(P_t + \mu_t - \sum_{\tau=t+L_G}^{T} \sigma_\tau + \sum_{\tau=t+L_G}^{T} \rho_\tau\right) = \frac{1}{a_t}\left(M - a_t P_t - b_t\left(\frac{\hat{W}_t}{X_t}\right)\right) \geq 0$$

Since sales $D_t > 0$, we have

$$\frac{1}{a_t}\left(M - a_t P_t - b_t(\frac{\hat{W}_t}{X_t})\right) > 0 \tag{3.69}$$

For the CF model, we have $K_1 > X_t$. This statement can be inferred from Fig. 3.1, where K_1 refers to the theoretical capacity indicated by the "fixed capacity" line and X_t is the output of the concave CF. Using this fact and Eq. (3.69) in (3.68), we have $\theta_t > 0$. From complementary slackness condition (3.51), if $\theta_t > 0$, $K_2 X_t + X_t \hat{W}_t - K_1 \hat{W}_t = 0$, implying that the capacity constraint is tight. *QED*.

Appendix 3.5: Proof of Proposition 2

From equation (3.44), we have $X_t \frac{\partial L}{\partial X_t} = 0$.

From $X_t > 0$, we have

$$\frac{\partial L}{\partial X_t} = \phi_t + \lambda_t - \pi_t + \theta_t(K_2 + \hat{W}_t) - \frac{b_t \hat{W}_t}{X_t^2}\left(P_t + \mu_t - \sum_{\tau=t+L_G}^{T} \sigma_\tau + \sum_{\tau=t+L_G}^{T} \rho_\tau\right) = 0$$

yielding

$$\pi_t = \phi_t + \lambda_t + \theta_t(K_2 + \hat{W}_t) - \frac{b_t \hat{W}_t}{X_t^2}\left(P_t + \mu_t - \sum_{\tau=t+L_G}^{T} \sigma_\tau + \sum_{\tau=t+L_G}^{T} \rho_\tau\right).$$

From equation (3.40), we have $R_t\ (c_t - \lambda_t) = 0$. From $R_t > 0$, we have $\lambda_t = c_{t.}$ Thus

$$\pi_t = c_t + \theta_t(K_2 + \hat{W}_t) - \frac{b_t \hat{W}_t}{X_t^2}\left(P_t + \mu_t - \sum_{\tau=t+L_G}^{T} \sigma_\tau + \sum_{\tau=t+L_G}^{T} \rho_\tau\right).$$

From (3.68), we have

$$\theta_t = \frac{1}{(K_1 - X_t)}\left(\chi_t + \frac{b_t}{X_t}\left(P_t + \mu_t - \sum_{\tau=t+L_G}^{T} \sigma_\tau + \sum_{\tau=t+L_G}^{T} \rho_\tau\right)\right)$$

$$\Rightarrow \frac{b_t}{X_t}\left(P_t + \mu_t - \sum_{\tau=t+L_G}^{T} \sigma_\tau + \sum_{\tau=t+L_G}^{T} \rho_\tau\right) = \theta_t(K_1 - X_t) - \chi_t$$

Using this relation in the expression for π_t, we have

$$\pi_t = \phi_t + c_t + \theta_t(K_2 + \hat{W}_t) + \frac{\hat{W}_t}{X_t}(\chi_t - \theta_t(K_1 - X_t)) \tag{3.70}$$

Since $\theta_t > 0$, we have $K_2 X_t + X_t \hat{W}_t - K_1 \hat{W}_t = 0$ from complementary slackness condition (3.51). Rewriting, we obtain $K_2 + \hat{W}_t = \frac{K_1 \hat{W}_t}{X_t}$. Using this in (3.70), we have

$$\pi_t = \phi_t + c_t + \frac{\theta_t K_1 \hat{W}_t}{X_t} + \frac{\hat{W}_t \chi_t}{X_t} - \frac{\theta_t K_1 \hat{W}_t}{X_t} + \hat{W}_t \theta_t = \phi_t + c_t + \frac{\hat{W}_t \chi_t}{X_t} + \hat{W}_t \theta_t > 0$$

QED.

Appendix 3.6: Proof of Proposition 3

We have two cases. Shipments can be made from the production quantity in the current period or from ending inventory from the previous period.

Case 1: $\hat{X}_t > 0$

In this case, sales are fulfilled from production in that period.

From equation (3.59),

$$\hat{X}_{t-L_G} \frac{\partial L}{\partial \hat{X}_{t-L_G}} = \hat{X}_{t-L_G}(c_{t-L_G} - \gamma_t + \delta_{t-L_G}) = 0$$

Since $\hat{X}_t > 0$ the expression in brackets equals zero. Rearranging the terms in the bracket, we obtain $\gamma_t = c_{t-L_G} + \delta_{t-L_G} > 0$, which implies that $\gamma_t > 0$ for any period with positive production.

Case 2: $\hat{I}_t > 0$
In this case, shipments take place from ending inventory of previous period.

From equation (3.57) for period $t-1$, we have

$$\gamma_t = h_{t-1} + \gamma_{t-1} \tag{3.71}$$

Let the last positive production have taken place t-τ periods before and sales in all subsequent periods be met from inventory resulting from this production. In other words, $\hat{X}_{t-L_G-\tau} > 0$ and $\hat{X}_{t-L_G-\tau+1} = \ldots = \hat{X}_{t-L_G} = 0$. Then from Case (1) we have $\gamma_{t\text{-}\tau} > 0$.

In addition, we have $\hat{I}_{t-\tau}, \hat{I}_{t-\tau+1}, \ldots, \hat{I}_{t-1} > 0$. Writing equation (3.71) for periods t-$\tau + 1$ to t, we have

$$\begin{aligned}
\gamma_{t-\tau+1} &= h_{t-\tau} + \gamma_{t-\tau}\\
\gamma_{t-\tau+2} &= h_{t-\tau+1} + \gamma_{t-\tau+1}\\
&\vdots\\
\gamma_t &= h_{t-1} + \gamma_{t-1}
\end{aligned}$$

Adding the above expressions, we get

$$\gamma_t = \sum_{i=t-\tau}^{t-1} h_i + \gamma_{t-\tau}$$

Since both terms on the right hand side are positive, we have $\gamma_t > 0$. *QED*.

Appendix 3.7

By Little's Law we have $L=\frac{\hat{W}_t}{X_t}$, implying that $\hat{W}_t = LX_t$. Noting that the capacity constraints will be tight, we have $X_t=\frac{K_1\hat{W}_t}{K_2+\hat{W}_t}$. The utilization u_t in period t can thus be calculated as $u_t=\frac{X_t}{K_1}=\frac{LX_t}{LX_t+K_2}$. Solving for K2, we obtain $K_2 = L(1-u_t)K_1$. Choosing $K_1 = M/2$, for $M = 1{,}000$ we obtain $K_1 = 500$. For $L = 1$ and $u_t = 0.8$, we obtain $K_2 = 100$.

Appendix 3.8

From WIP Balance constraint of CF model, we have:

$$W_t = W_{t-1} - X_t + R_t$$

Writing the constraint for WIP level as an equality, we have $\hat{W}_t=\frac{W_{t-1}+W_t}{2}$, implying $W_t=2\hat{W}_t-W_{t-1}$. Setting the two expressions for W$_t$ equal to each other, we find that $X_t=R_t-2\hat{W}_t+2W_{t-1}$

Writing the CF constraint as an equality,

$$X_t = f(\hat{W}_t) = \frac{K_1\hat{W}_t}{K_2+\hat{W}_t}$$

Comparing both equations for X_t, we have

$$R_t - 2\hat{W}_t + 2W_{t-1} = \frac{K_1\hat{W}_t}{K_2+\hat{W}_t}$$

Solving the resulting quadratic in $\hat{W}_t$

$$\hat{W}_t = \frac{(R_t-2K_2+2W_{t-1}-K_1)\pm\sqrt{(R_t-2K_2+2W_{t-1}-K_1)^2+8K_2(R_t+W_{t-1})}}{4}$$

We use the positive root for calculating WIP level in a period when simulating FLT decisions under congestion.

References

Adida, E., & Perakis, G. (2006). A robust optimization approach to dynamic pricing and inventory control with no backorders. *Mathematical Programming Series B, 107,* 97–129.

Adida, E., & Perakis, G. (2010). Dynamic pricing and inventory control: robust vs. stochastic uncertainty models: A computational study. *Annals of Operations Research, 181,* 125–157.

Agnew, C. (1976). Dynamic modeling and control of some congestion prone systems. *Operations Research, 24*(3), 400–419.

Ahn, H., Gumus, M., & Kaminsky, P. (2007). Pricing and manufacturing decisions when demand is a function of prices in multiple periods. *Operations Research, 55*(6), 1039–1057.

Akcali, E., Nemoto, K., & Uzsoy, R. (2000). Quantifying the benefits of cycle-time reduction in semiconductor wafer fabrication. *IEEE Transactions on Electronics Packaging Manufacturing, 23,* 39–47.

Allison, R. A. H., Yu, J., Tsai, L. H., Liu, C., Drummond, M., Kayton, D., Sustae, T., & Witte, J. (1997). Macro model development as a bridge between factory level simulation and LP enterprise systems. *IEEE/CPMT International Electronics Manufacturing Technology Symposium*: 408–416.

Asmundsson, J. M., Rardin, R. L., Turkseven, C. H., & Uzsoy, R. (2009). Production planning models with resources subject to congestion. *Naval Research Logistics, 56,* 142–157.

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

Boyaci, T., & Ray, S. (2003). Product differentiation and capacity cost interaction in time and price sensitive markets. *Manufacturing and Service Operations Management, 5*(1), 18–36.

Buzacott, J. A., & Shanthikumar, J. G. (1993). *Stochastic models of manufacturing systems.* Englewood Cliffs, NJ, Prentice-Hall.

Charnsirisakskul, K., Griffin, P., & Keskinocak, P. (2004). Order selection and scheduling with leadtime flexibility. *IIE Transactions, 36,* 697–707.

Charnsirisakskul, K., Griffin, P., & Keskinocak, P. (2006). Pricing and scheduling decisions with leadtime flexibility. *European Journal of Operational Research, 171,* 153–169.

Chatterjee, S., Slotnick, S. A., & Sobel, M. J. (2002). Delivery guarantees and the interdependence of marketing and operations. *Production and Operations Management, 11*(3), 393–410.

Chen, Z. L., & Hall, N. G. (2010). The coordination of pricing and scheduling decisions. *Manufacturing and Service Operations Management, 12*(1), 77–92.

de Kok, A. G., & Fransoo, J. C. (2003). *Planning supply chain operations: definition and comparison of planning concepts. OR Handbook on supply chain management*. A. G. de Kok & S. C. Graves (597–675). Amsterdam: Elsevier.

Dellaert, N. P. (1991). Due date setting and production control. *International Journal of Production Economics, 23,* 59–67.

Deng, S., & Yano, C. A. (2006). Joint production and pricing decisions with setup costs and capacity constraints. *Management Science, 52,* 741–756.

Donohue, K. L. (1994). The economics of capacity and marketing measures in a simple manufacturing environment. *Production and Operations Management, 3*(2), 78–99.

Duenyas, I. (1995). Single facility due date setting with multiple customer classes. *Management Science, 41*(4), 608–619.

Duenyas, I., & Hopp, W. J. (1995). Quoting customer lead times. *Management Science, 41,* 608–619.

Easton, F. F., & Moodie, D. R. (1999). Pricing and lead time decisions for make-to-order firms with contingent orders. *European Journal of Operational Research, 116,* 305–318.

Elhafsi, M. (2000). An operational decision model for lead-time and price quotation in congested manufacturing systems. *European Journal of Operational Research, 126,* 355–370.

Elhafsi, M., & Rolland, E. (1999). Negotiating price/delivery date in a stochastic manufacturimg environment. *IIE Transactions, 31,* 255–270.

Eliashberg, J., & Steinberg, R. (1991). Marketing-production joint decision-making. *Management science in marketing, handbooks in operations research and management science*. J. Eliashberg and J. D. Lilien, North Holland: 827–880.

Elmaghraby, W., & Keskinocak, P. (2003). Dynamic pricing in the presence of inventory considerations: Research overview, current practices and future directions. *Management Science, 49*(10), 1287–1309.

Graves, S. C. (1986). A tactical planning model for a job shop. *Operations Research, 34,* 552–533.

Hackman, S. T., & Leachman, R. C. (1989). A general framework for modeling production. *Management Science, 35,* 478–495.

Hopp, W. J., & Spearman, M. L. (2001). *Factory physics: Foundations of manufacturing management*. Boston, Irwin/McGraw-Hill.

Johnson, L. A., & Montgomery, D. C. (1974). *Operations research in production planning, scheduling and inventory control*. New York: John Wiley.

Kacar, N. B., & Uzsoy, R. (2010). Estimating clearing functions from simulation data. *Winter Simulation Conference*. B. Johansson, Jain, S., Montoya-Torres, J., Hugan, J., Yucesan, E. Baltimore, MD.

Karmarkar, U. S. (1989). Capacity loading and release planning with work-in-progress (WIP) and lead-times. *Journal of Manufacturing and Operations Management, 2*(105-123).

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

Keskinocak, P., & Tayur, S. (2004). Due-date management policies. In D. Simchi-Levi, S. D. Wu, & Z. M. Shen (Eds.), *Supply chain analysis in the e-business era: Handbook of quantitative supply chain analysis*. Kluwer Academic Publishers.

Leachman, R. C., & Ding, S. (2007). Integration of speed economics into decision-making for manufacturing management. *International Journal of Production Economics, 107,* 39–55.

Liu, L. M., Parlar, M., & Zhu, S. X. (2007). Pricing and lead time decisions in decentralized supply chains. *Management Science, 53*(5), 713–725.

Low, D. W. (1974). Optimal dynamic pricing policies for an M/M/s queue. *Operations Research, 22,* 545–561.

Medhi, J. (1991). *Stochastic models in queuing theory*. Academic Press.

Missbauer, H. (2009). Models of the transient behaviour of production units to optimize the aggregate material flow. *International Journal of Production Economics, 118*(2), 387–397.

Missbauer, H., & Uzsoy, R. (2010). *Optimization models for production planning. Planning production and inventories in the extended enterprise: A state of the art handbook*. K. G. Kempf, P. Keskinocak and R. Uzsoy (437–508). New York: Springer.

Orcun, S., Uzsoy, R., & Kempf, K. G. (2006). Using system dynamics simulations to compare capacity models for production planning. *Winter Simulation Conference*. Monterey, CA.

Pahl, J., Voss, S., & Woodruff, D. L. (2005). Production planning with load dependent lead times. *4OR: A Quarterly Journal of Operations Research, 3,* 257–302.

Pahl, J., Voss, S., & Woodruff, D. L. (2007). Production planning with load dependent lead times: An update of research. *Annals of Operations Research, 153,* 297–345.

Palaka, K., Erlebacher, S., & Kropp, D. H. (1998). Lead-time setting capacity utilization, and pricing decisions under lead-time dependent demand. *IIE Transactions, 30,* 151–163.

Pekgun, P., Griffin, P. M., & Keskinocak, P. (2008). Coordination of marketing and production for price and leadtime decisions. *IIE Transactions, 40*(1), 12–30.

Plambeck, E. L. (2004). Optimal leadtime differentiation via diffusion approximation. *Operations Research, 52*(2), 213–228.

Ray, S., & Jewkes, E. M. (2004). Customer lead time management when both demand and price are lead time sensitive. *European Journal of Operational Research, 153,* 769–781.

Selçuk, B., Fransoo, J. C., & de Kok, A. G. (2007). Work in process clearing in supply chain operations planning. *IIE Transactions, 40,* 206–220.

So, K. C., & Song, J.-S. (1998). Price, delivery time guarantees and capacity selection. *European Journal of Operational Research, 111,* 28–49.

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

Spitter, J. M., A. G. de Kok and N. P. Dellaert (2005a). Timing production in LP models in a rolling schedule. *International Journal of Production Economics, 93–94,* 319–329.

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

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

Sterman, J. D. (2000). *Business dynamics: Systems thinking and modeling for a complex world.* New York: McGraw-Hill.

Swann, J. L. (2001). Dynamic pricing models to improve supply chain performance. *Department of Industrial Engineering and Management Sciences*. Evanston, IL 60601, Northwestern University.

Tardif, V., & Spearman, M. L. (1997). Diagnostic scheduling in finite-capacity production environments. *Computers and Industrial Engineering, 32,* 867–878.

Upasani, A., & Uzsoy, R. (2008). Incorporating manufacturing lead times in joint production-marketing models: A review and further directions. *Annals of Operations Research, 161,* 171–188.

Webster, S. (2002). Dynamic pricing and lead time policies for make to order systems. *Decision Sciences, 33*(4), 579–599.

Yano, C. A., & Gilbert, S. M. (2003). Coordinated pricing and production/procurement decisions: A review. *Managing business interfaces: Marketing, engineering and manufacturing perspectives*. A. Charkarvarty and J. Eliashberg, Kluwer Academic Publishers: 65–103.