Using a mathematical programming model to examine the marginal price of capacitated resources
Ali Kefeli, Reha Uzsoy*, Yahya Fathi, Michael Kay
Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Campus Box 7906, Raleigh, NC 27695-7906, USA

abstract
Accurate information on dual prices of capacitated resources is of interest in a number of applications, such as cost allocation and pricing. To gain insight we focus on the dual prices of capacity and demand in a single-stage single-product production-inventory system, and discuss their interpretation. In particular, we examine the behavior of two different production planning models: a conventional linear programming model and a nonlinear model that captures queuing behavior at resources in an aggregate manner using nonlinear clearing functions. The classical linear programming formulation consistently underestimates the dual price of capacity due to its failure to capture the effects of queuing. The clearing function formulation, in contrast, produces positive dual prices even when utilization is below one and exhibits more realistic behavior, such as holding finished inventory at utilization levels below one.

1. Introduction
While linear programming (LP) models have been used extensively for production planning problems since the 1950s (e.g., 1974; Missbauer and Uzsoy, in press), their limitations have been increasingly discussed in recent years. Since these approaches do not explicitly model work-in-process (WIP) inventories, as opposed to finished intermediate inventories between stations, they assume that resources can operate at full utilization while maintaining fixed cycle times, which contradicts both empirical observations and predictions of queuing models. Another anomaly is that a resource can have a nonzero dual price only when the associated capacity constraint is tight or, in queuing terminology, utilization of that resource reaches one. However, it is well known from industrial experience and queuing models that manufacturing performance measures such as the expected WIP and cycle time increase nonlinearly with utilization. This suggests that with an appropriate objective function, additional capacity can improve system performance even when utilization is less than one, implying a nonzero dual price for the resource at lower utilization levels. Since most production systems are governed to a large degree by queuing behavior, this inconsistency between the LP models used for capacity allocation among products and the queuing models used to describe steady-state behavior poses an interesting research direction that has not been explored extensively.

In this paper we examine the behavior of an alternative mathematical programming model based on nonlinear clearing functions (CFs) suggested by Karmarkar (1989) in a simple single-stage single-product production-inventory system. While models of this type have been successfully applied to multi-product multi-stage systems (e.g., Asmundsson et al., 2006, 2009), we focus on a single-stage single-product system in order to maintain the paper’s focus on insights into the different dual prices obtained by the two models.

Our results show that the CF model explicitly captures the nonlinear relationships between performance measures and system workload, providing a more accurate representation of actual capabilities of the production system. We derive closed-form expressions for the dual prices of capacity and demand and illustrate qualitative differences in both production plans and dual prices between those from the CF model and the conventional LP model with numerical examples. We conclude the paper with a summary and some directions for future work.

2. Previous related work
The use of mathematical programming models for production planning has a long history, dating back to the seminal work of Modigliani and Hohn (1955). The majority of these are LP formulations, where the planning horizon is divided into discrete time periods and two principal sets of constraints are used to
represent the operation of the production system. One set of constraints represents conservation of material flow into and out of finished goods inventories (FGIs) at the end of each planning period, assuming that production and demand rates are uniform within each period. A second family of constraints limits the total amount of work a resource, such as a machine, can perform in a period. A wide range of such models has been presented in the literature over the years (e.g., Hackman and Leachman, 1989; Johnson and Montgomery, 1974; Voss and Woodruff, 2003).

However, these LP models have a number of drawbacks. Since they assume all production rates are constant in a planning period, they do not model WIP at all, and do not assign it a cost in the objective function. Hence they allow resources to change their production rates with complete flexibility as long as utilization is less than one. For the same reason, when backlogging is not allowed, these models will hold FGI only when demand exceeds resource capacity in some period. Although queuing models and industrial experience suggest that when capacity is fully utilized, very high cycle times and WIP levels will result, the LP models assume that resources can maintain a fixed cycle time independent of utilization. Finally, they will only permit nonzero dual prices only for capacity constraints when that constraint is tight at optimality, implying a utilization level of one for that resource during that period.

The marginal or dual price of a production resource with limited capacity, which represents the potential change in objective function as additional capacity becomes available, is of interest in several situations. One example is in short-term scheduling and dispatching, where a number of authors have proposed dispatching rules based on cost–benefit ratios, where the value of processing a job at a given time is contrasted with the marginal price of the resource consumed by scheduling it at that time (e.g., Morton et al., 1988). Yet another related area is in the estimation of setup costs. It can be argued that many of the costs related to setup changes, such as manpower, are fixed relative to the decision whether or not to set up at a particular point in time. In many cases, the principal cost of the setup is the lost revenue due to output not being produced during the setup. If utilization is low and demand can still be met when the setup is performed, the lost revenue, and thus cost of the setup, will be zero. However, if utilization is high the opportunity cost of a long setup can be quite substantial. Another potential use of the marginal price of capacity has been suggested in the accounting literature (e.g., Kaplan and Thompson, 1971), where overhead costs are allocated among products based on their consumption of capacity at key resources.

The elegant duality theory associated with LPs (e.g., Bazaraa et al., 2004) directly yields the marginal price of a capacitated resource in a planning period as the optimal value of the dual variable associated with that capacity constraint. However, by the LP optimality conditions, a dual variable will only take a nonzero dual value when the associated capacity constraint is tight at optimality, i.e. when the resource utilization is equal to one. This suggests that there is no benefit to the objective function from adding capacity to resources that are not fully utilized. However, this contradicts observations from industrial practice, simulation models, and queuing models of manufacturing systems (e.g., Buzacott and Shanthikumar, 1993; Hopp and Spearman, 2001). All these suggest that performance measures of a production system, especially WIP levels and cycle times, increase nonlinearly with resource utilization, and that significant increases in these quantities can be observed even at utilization levels substantially below one. Since many manufacturing resources such as machines exhibit queuing behavior, this inconsistency between the production planning models used for capacity allocation over time and the queuing models representing the behavior of resources is troubling.

The queuing results would appear to suggest that even at utilization levels below one, adding capacity to the system ought to result in improved performance due to reduced WIP levels and cycle times under at least some conditions. This, in turn, implies that we ought to observe nonzero dual prices on capacity constraints at utilization levels less than one. A number of authors (Graves, 1986; Karmarkar, 1989; Srinivasan et al., 1988) have proposed mathematical programming models for resource allocation that exhibit such behavior.

Several alternative approaches to estimating the dual prices of production resources have been proposed in the literature. Morton et al. (1988) propose a dispatching rule for job shop scheduling based on a cost–benefit ratio, where the cost of a unit of machine time at a particular point in time is estimated as the increase in tardiness penalties caused by a differential delay in machine availability. The latter estimates are obtained by simulating the shop to the end of the current busy period, the time at which the machine will be idle. Roundy et al. (1991) propose an alternative approach, where an aggregate planning model is used to estimate costs of performing a particular operation at a particular point in time; the dispatching module then dispatches jobs based on these costs. The costs are estimated as values of the Lagrange multipliers associated with machine capacity constraints in each time interval. Hoitomt et al. (1993) present a similar formulation for job shop scheduling, where time intervals are reduced in length, giving a dual price for each individual time unit. These approaches focus on estimating the dual price of capacity, noting that the dual price of a resource will depend on the state of the system at that point in time.

Another approach to estimating the dual price of capacity has been through queuing models, yielding an average price over a long period in which the system is in steady state. Banker et al. (1986) use an $M/G/1$ queuing model of a production system with lot sizing to examine the effects of adding new products to an existing facility, focusing on the costs of holding WIP. They study the marginal benefits of adding capacity and show that it is beneficial to add capacity at all levels of utilization, but that the benefit decreases as utilization decreases. This, in our context, corresponds to a dual price of capacity that increases with utilization. Morton and Singh (1988) approach the same problem using a variation of the busy period methodology from Morton et al. (1988), and show that under steady state assumptions they obtain the same dual prices for capacity as Banker et al. (1986).

It is important to bear in mind that the nature of a dual price will depend on the objective function or performance measure of the model from which it is derived. The short-term dispatching and scheduling approaches outlined above, as well as the steady-state queuing models, do not consider the effects of FGI that is held due to insufficient capacity to meet demand in a specific time period. The LP models of production planning, on the other hand, explicitly consider the possibility of meeting future demand by building FGI in earlier periods, but do not consider the effects of resource utilization on queue lengths within the production system. Clearly both types of costs, those due to WIP accumulating in the system as well as FGI that is held to address temporary capacity shortfalls, are relevant, and hence a model that integrates both would appear to be desirable. In this paper we explore this issue for a single-product single-stage production-inventory system, which has not been done before to the best of our knowledge.

3. Classical LP formulation

In this section we present the classical LP production planning model we use as a benchmark. For simplicity of exposition and without loss of generality, we shall assume that all production
4. Clearing function formulation

CFs, proposed by Graves (1986), Karmarkar (1989), and Srinivasan et al. (1988), express the expected throughput of a capacitated resource over a given period of time as a function of workload of the resource over that period, which, in turn, is defined by the amount of work available for the resource to process in that period. For now we shall use the term “WIP” and the generic variable $W$ to denote any reasonable measure of the WIP inventory level, whether carried over from a previous period or released during the current one.

To motivate the use of a nonlinear CF, it is helpful to begin with a single resource that can be modelled as a $G/G/1$ queuing system in steady state. The expected number in system, or, equivalently, the expected WIP, for a single server is given by Medhi (1991) as

$$W = \frac{(c^2 + c_2)}{2} \frac{\rho^2}{(1-\rho)} + \rho$$

where $c_2$ and $c_1$ denote the coefficients of variation of interarrival and service times, respectively, and $\rho$ is the utilization of the server. Setting $c = (c^2 + c_2)/2$ and rearranging (9) we obtain a quadratic in $W$, whose positive root yields the desired $\rho$ value. Solving for $\rho$ with $c > 1$, we obtain

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

which has the desired concave form. When $0 \leq c < 1$, the other root of the quadratic will always give positive values for $\rho$. When $c=1$, (9) simplifies to yield $\rho = W/(1+W)$ again of the desired concave form. If we use utilization as a surrogate for output, we see that for a fixed $c$ value, utilization, and hence throughput, increases with WIP but at a declining rate. Utilization, and hence output, decreases with $c$ 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 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 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. In addition to the queuing approaches outlined above, clearing functions can also be estimated using industrial data, or data obtained from a simulation model of the system under study, as suggested by Karmarkar (1989), Asmundsson et al. (2006), and Srinivasan et al. (1988). These methods involve collecting observations of average resource workload and throughput and then fitting a clearing function to these data.

Different methods of generating CFs for optimization models have different advantages and disadvantages. A CF estimated by a steady-state queuing model assumes at least approximate steady state behavior of the queue being modelled in a planning period, which may be problematic in systems with long cycle times at individual resources. On the other hand, while CF estimation through simulation and industrial data do not require steady state assumptions, they are only as valid as the simulation model or data they are derived from and will require updating as products or processes change. The development of methods for deriving and validating non-steady state clearing functions remains an active area of research (Asmundsson et al., 2009; Selcuk et al., 2007; Riaño, 2003). An extensive discussion of the issues related to clearing functions and their use in production planning models is given by Pahl et al. (2005) and Missbauer and Uzsoy (in press). In this paper
we shall assume that an appropriate clearing function can be obtained using techniques suggested in these references, and focus on the behavior of the resulting planning models.

Our starting point is the single-product single-stage model of Karmarkar (1989). Let us define the following decision variables:

\[ X_t = \text{number of units of the product produced in period } t, \]
\[ R_t = \text{number of units of the product released into the stage at the beginning of period } t, \]
\[ W_t = \text{number of units of the product in work in process (WIP) at the end of period } t, \]
\[ I_t = \text{number of units of the product in finished goods inventory (FGI) at the end of period } t. \]

Let \( f(.) \) denote the clearing function and \( D_t \) the demand for the product (in units) in period \( t \). Then Karmarkar’s formulation is as follows:

\[
\text{minimize } \sum_{t=1}^{T} \left( r_t R_t + w_t W_t + p_t X_t + h_t I_t \right) \quad (11) \\
\text{subject to } \\
W_t = W_{t-1} + R_t - X_t \quad \forall t \quad (12) \\
I_t = I_{t-1} + X_t - D_t \quad \forall t \quad (13) \\
X_t \leq f(W_{t-1} + R_t) \quad \forall t \quad (14) \\
R_t, W_t, X_t, I_t \geq 0 \quad \forall t \quad (15)
\]

where \( r_t, w_t, p_t, \) and \( h_t \) denote the unit cost coefficients of raw material releases, WIP holding, production, and FGI holding, respectively, and \( T \) is the final period in the planning horizon. We follow Karmarkar in writing our CF as a function of the sum \( W_{t-1} + R_t \), the resource load for period \( t \), or, in other words, the total amount of work that becomes available for processing during the period. Since the formulation distinguishes between WIP and FGI, flow conservation constraints (12) and (13) are required for both quantities, which serve distinct purposes. In periods when demand cannot be satisfied due to limited capacity, it can be met with FGI carried over from previous periods. WIP, on the other hand, accumulates in front of the resources as jobs in queue and in process. This approach differs from traditional LP models since it links the expected throughput of the resource in period \( t \) to the resource load (incoming WIP and new releases) in that period.

The solution to the CF formulation provides the decision maker with a time-phased release plan, i.e. how many new jobs to release to the production resource in each period, specified by the decision variables \( R_t \). We assume that job releases are under complete control of the decision maker and that raw material availability is not a constraint. This allows us to study the dual prices of capacity in isolation from the effects of external factors, allowing more intuitive results to be obtained. Another contrast to the LP formulations such as those of Hackman and Leachman (1989) is that lead times do not appear in the formulation; releases and lead times are jointly optimized, allowing lead times to vary over the planning horizon.

At this point, in order to obtain a more complete characterization of the dual prices, we introduce an additional set of decision variables to our formulation, motivated by the multi-product CF model of Asmundsson et al. (2006). We define

\[ Z_t = \text{fraction of the maximum output in period } t, \]
and modify (14) to include these decision variables as follows:

\[ X_t \leq Z_t f(W_{t-1} + R_t) \quad \forall t \quad (16) \]

In a multi-product formulation, these variables are defined for each product and period to allocate the available output of a production resource among different products that use it. By definition, the summation of these variables over all products for a given resource should be equal to 1. Since we have a single product, we add one additional constraint to our formulation:

\[ Z_t = 1 \quad \forall t \quad (17) \]

Although these additional variables and constraints may seem redundant, they help us to better characterize the dual prices of capacity in the CF formulation. Following Asmundsson et al. (2006) and Missbauer (2002), we approximate the CF \( f(.) \) using outer linearization and incorporate (16) and (17) into our formulation, yielding the following LP:

\[
\text{minimize } \sum_{t=1}^{T} \left( r_t R_t + w_t W_t + p_t X_t + h_t I_t \right) \quad (18) \\
\text{subject to } \\
W_t = W_{t-1} + R_t - X_t \quad \forall t \quad (19) \\
I_t = I_{t-1} + X_t - D_t \quad \forall t \quad (20) \\
X_t \leq Z_t f(W_{t-1} + R_t) + Z_t \beta_t \quad \forall t \quad (21) \\
Z_t = 1 \quad \forall t \quad (22) \\
R_t, W_t, X_t, I_t, Z_t \geq 0 \quad \forall t \quad (23)
\]

where \( \beta_t \) is the intercept of segment \( c \) of the piecewise linearized CF. The first linear segment \( (c=1) \) will have a slope of one \( (\beta_1 = 1) \) since at low utilization all work in the system can be completed within a planning period, and an intercept of zero \( (\beta_0 = 0) \), since production cannot take place without some WIP being present. The last linear segment \( (c=n) \) will have a slope of zero \( (\beta_n = 0) \) since increasing the load of the resource will not increase output. To capture the concavity of the clearing function, successive segments will have strictly increasing intercepts \( (\beta_1 < \beta_2 < \beta_3 < \cdots < \beta_n) \) and strictly decreasing slopes \( (\beta_1 > \beta_2 > \beta_3 > \cdots > \beta_n) \) as seen in Fig. 1.

Eliminating the \( W_t \) and \( I_t \) variables in a similar fashion to that used for the LP model in the previous section, making appropriate substitutions in the objective function and capacity constraints, and defining boundary conditions, we obtain the following formulation that will be the basis of our analysis:

\[
\text{minimize } \sum_{t=1}^{T} \left[ \left( r_t + \sum_{t=1}^{T} p_t \right) R_t + \left( p_t + \sum_{t=1}^{T} h_t \right) Z_t \right] X_t \quad (24) \\
\text{subject to } \\
\sum_{t=1}^{T} X_t \geq \sum_{t=1}^{T} D_t - L_0 \quad \forall t \quad (25) \\
- \sum_{t=1}^{T} X_t + \sum_{t=1}^{T} R_t \geq -W_0 \quad \forall t \quad (26)
\]

\[
\text{Fig. 1. A typical CF and outer linearized segments.}
\]
where the dual prices associated with each constraint set are given in parentheses. We now proceed to analyze the dual of this formulation.

5. Analysis of the CF formulation

The dual of this formulation (24)–(30) is given by

$$\text{minimize} \quad \sum_{t=1}^{T} \left[ \left( \sum_{c=1}^{C} D_t - I_0 \right) \gamma_t - W_0 \Gamma_t - \sum_{c=1}^{C} a_c W_0 \sigma_{ct} + \lambda_t \right]$$

subject to

$$\sum_{t=1}^{T} \gamma_t - \sum_{t=1}^{T} \Gamma_t - \sum_{c=1}^{C} \sigma_{ct} - \sum_{t=1}^{T} \sum_{c=1}^{C} a_c \sigma_{ct} \leq p_t$$

$$+ \sum_{t=1}^{T} (h_t - w_t) \quad \forall t \neq T \quad (X_t)$$

$$\gamma_t - \Gamma_t - \sum_{c=1}^{C} \sigma_{ct} \leq p_t + h_t - w_t \quad (X_t)$$

$$\sum_{t=1}^{T} \Gamma_t + \sum_{t=1}^{T} \sum_{c=1}^{C} a_c \sigma_{ct} \leq r_t + \sum_{t=1}^{T} w_t \quad \forall t \quad (R_t)$$

$$\lambda_t + \beta^2 \sigma_{ct} \leq 0 \quad \forall t \quad (Z_t)$$

$$\gamma_t, \Gamma_t, \sigma_{ct} \geq 0 \quad \forall t$$

$$\lambda_t \text{ free} \quad \forall t$$

Note that any production plan will consist of congested periods where $W_t > 0$, non-congested periods where $W_t = 0$ and $X_t > 0$, and idle periods where $W_t = 0$ and $X_t = 0$. Since there is no work release, production, or WIP in idle periods, they are not of great interest. In non-congested periods, production takes place but no WIP is carried from one period to the other. This is possible if the system operates at low utilization such that all work released into the system is processed in the same period, implying that the first segment ($c = 1$) of the CF is tight (see Fig. 1). In congested periods, on the other hand, all work available to the resource cannot be completed in one period due to congestion, and some is carried over to the next period as WIP. This means the system is operating at higher utilization and a segment of the CF with index $c > 1$ is tight. (In certain extreme cases, where demand falls very sharply in a short period of time, it is possible to have a congested period $t$ with $X_t = 0$ due to the presence of high WIP levels that was needed to meet high demand in earlier periods but for which demand no longer exists. In the remainder of this paper we shall assume demand is sufficiently regular to avoid this extreme behavior.)

We will assume $W_0 = 0$ throughout our analysis since feasibility can be ensured by an appropriate $I_0$. As stated earlier, our objective is to explore the behavior of dual prices in the context of this LP model. To this end, we first examine the behavior of the system in a congested interval, defined as a collection of consecutive congested periods, starting with a period $s$ and ending with a period $s' > s$. We denote such a congested interval by $K = \{s, s+1, s+2, \ldots, s'\}$ such that $W_{t-1} = 0, W_t > 0$ for all $t \in K$, and $W_{s'} = 0$.

We first note, by the complementary slackness condition of LP optimality, that $W_0 > 0$ implies that $\gamma_0 = 0$ for all $t \in K$ in the optimal dual solution. Assume that for all $t \in K$ we have $R_t > 0$, which in turn implies that $X_t > 0$, i.e. if we release any work in period $t$, there must be some production in that period, since otherwise we can release the work in a later period and save the WIP holding cost. Under these assumptions, constraints (32) and

$$\sum_{t=1}^{T} \gamma_t - \sum_{t=1}^{T} \Gamma_t - \sum_{c=1}^{C} \sigma_{ct} - \sum_{t=1}^{T} \sum_{c=1}^{C} a_c \sigma_{ct}$$

$$= p_t + \sum_{t=1}^{T} (h_t - w_t) \quad \forall t \in K$$

(38)

If $t \in K$ the corresponding expression for $t = T$ is

$$\gamma_T - \sum_{c=1}^{C} \sigma_{ct} \leq p_T + h_T - w_T$$

(40)

We now use these relationships to derive expressions for the optimal values of $\sigma_{ct}, \gamma_t$, and $\lambda_t$ in a congested interval. For brevity let $\Phi_t = \sum_c a_c \sigma_{ct}$ and rewrite (39) as

$$\sum_{t=1}^{T} \gamma_t = \sum_{t=1}^{T} \Gamma_t + \sum_{c=1}^{C} \Phi_t = r_t + \sum_{t=1}^{T} w_t \quad \forall t \in K$$

(41)

Writing (41) for successive periods from $s'$ to $s$ and solving recursively yields

$$\Phi_t = (\rho_t - \rho_{t+1}) + \sum_{c=1}^{C} (\sigma_{ct} - \sigma_{ct+1}) + \sum_{t=1}^{T} w_t \quad \forall t \in K \{s'\}$$

Under time stationary costs ($\rho_t = \rho, w_t = w, h_t = h,$ and $\rho_t = p \forall t$) this simplifies to

$$\Phi_t = w \quad \forall t \in K \{s'\}$$

(42)

To derive the values of $\gamma_t$, we return to (38) with the additional assumption that $R_{t+1} > 0$. We can then write (39) for period $t + 1$ and substitute in (38) to obtain

$$\sum_{t=1}^{T} \gamma_t - \sum_{t=1}^{T} \Gamma_t - \sum_{c=1}^{C} \sigma_{ct} \left( \rho_{t+1} + \sum_{t=1}^{T} w_t - \sum_{t=1}^{T} \Gamma_t \right)$$

$$= p_t + \sum_{t=1}^{T} (h_t - w_t) \quad \forall t \in K$$

(43)

Following the same recursive approach for (43) as we used for (41) we obtain

$$\gamma_t = h_t + \left( \sum_{c=1}^{C} (\sigma_{ct} - \sigma_{ct+1}) \right) + (\rho_t - \rho_{t+1}) - (w_t - w_{t+1})$$

$$+ (\rho_t + \rho_{t+1} - 2\rho_t) \quad \forall t \in K \{s'\}$$

which under time stationary costs simplifies to

$$\gamma_t = h + \left( \sum_{c=1}^{C} (\sigma_{ct} - \sigma_{ct+1}) \right) \quad \forall t \in K \{s'\}$$

(44)

We now concentrate on (42) and (44). To obtain more general forms of these relationships, we note that in an optimal solution, for a given period $t$, no more than two segments of the CF can be tight. For the time being, without loss of generality, let two segments $c^*(t) < c^*(t+1)$ of the CF be tight in the optimal solution for a given period $t$. By complementary slackness, the dual prices
of segments \( c \) that are not tight in an optimal solution in period \( t \) must be equal to zero, i.e. \( \sigma_{ct} = 0 \quad \forall c \neq c^*(t) \) and \( c \neq c^*(t) + 1 \). Hence

\[
\phi_t = \sum_c \sigma_{ct} = \lambda_t + \sigma_{c^*(t),t} + \sigma_{c^*(t) + 1,t}
\]

\[
\sum_c \sigma_{ct} = \sigma_{c^*(t),t} + \sigma_{c^*(t) + 1,t}
\]

If we also assume that there are two tight segments of the CF in period \( t+1 \) we obtain the following general forms of (42) and (44):

\[
\phi_t = w = \lambda_t + \sigma_{c^*(t),t} + \sigma_{c^*(t) + 1,t}
\]

\[
\gamma_t = h = \left( \sigma_{c^*(t),t} + \sigma_{c^*(t) + 1,t} \right) \left( \sigma_{c^*(t) + 1,t} \right) \quad t \in K^\prime \quad (48)
\]

Recall that (47) and (48) both hold for a period \( t \) such that \( t \in K^\prime \), i.e. with \( W_t > 0 \) by definition and with the assumption \( R_t > 0 \), which implies \( X_t > 0 \). Expression (48) also requires \( R_{t+1} > 0 \).

If we further assume positive FGI carried over from period \( t \) to period \( t+1 \), i.e. \( l_t > 0 \), by complementary slackness applied to (25), we will have \( \gamma_t = 0 \). Thus (48) will reduce to

\[
h = \lambda_t + \sigma_{c^*(t) + 1,t} \quad t \in K^\prime \quad (49)
\]

This implies that as long as work is released and FGI is carried in consecutive periods in a congested interval, the sum of the nonzero dual prices of capacity increases by \( h \) for each consecutive time period the system holds FGI.

It is clear that \( \sigma_{ct} \) is related to the dual price of capacity. However, the fact that two segments of the CF may be tight in a given period renders an intuitive expression hard to obtain. We now argue that the dual variables \( \lambda_t \) are a more informative estimate of the dual price of capacity in our CF formulation. An optimal solution to our CF formulation will have at most two segments of the CF tight, yielding two different \( \sigma_{ct} \) values for a given period. While these values are governed by (42), they do not provide a single number that can be used as the dual price of capacity. However, the extra decision variables \( Z_t \) provide a single number that can be used as the dual price of capacity. An estimate of the dual price of capacity in our CF formulation is therefore given by

\[
\gamma_t = h = \left( \sigma_{c^*(t),t} + \sigma_{c^*(t) + 1,t} \right) \left( \sigma_{c^*(t) + 1,t} \right) \quad t \in K^\prime
\]

This clearly reflects the fact that as long as work is released and FGI is carried in consecutive periods in a congested interval, the sum of the nonzero dual prices of capacity increases by \( h \) for each consecutive time period the system holds FGI.

6. Non-congested periods

So far, we have studied congested periods, where we have \( W_t > 0 \) and \( l_t > 0 \). For completeness, we now consider non-congested periods (see Table 1), where the system works just in time—utilization is low, so work is released and processed into output in the same period without WIP being carried between periods.

Using (21) for simplicity of exposition, we consider a non-congested period with demand \( d_t \). Since backlogging is not allowed, \( X_t = d_t \). By the definition of a non-congested period, only the first segment of the CF is tight, yielding

\[ X_t = W_{t-1} + R_t \quad \forall t \in K \quad (51) \]

and an optimal solution will give \( W_t = W_{t-1} = l_t = 0 \) and \( R_t = R_{t+1} = X_t = d_t \). Hence this is a highly degenerate solution, where all four constraints (25), (26), (27), and (29) are tight. In this situation the critical binding constraint is (25), the demand constraint, since until demand increases to the point that the second segment of the CF becomes tight, there is no benefit to additional capacity or releases. This fact is clearly reflected in (50)—since in a non-congested period only first segment of the CF is tight we have \( \sigma_{ct} = 0 \), \( \forall c \neq 1 \) and \( \beta_i = 0 \). This clearly implies \( \gamma_t = 0 \). Therefore, the non-congested periods do not yield any interesting insight on the dual price of capacity, following the behavior of the LP model when demand is below capacity and no inventory is required in the future.

7. Numerical examples

In order to illustrate the differences in the dual information obtained from the LP and the CF models under different conditions, we examine a numerical example with one product, one resource, and a planning horizon of \( T = 50 \) periods. We consider the demand scenario shown in Fig. 2, where demand remains below the theoretical capacity in the first 30 periods, and then increases to exceed capacity several times in the latter part increases. Additional capacity becomes more and more desirable, i.e. a higher dual price, as we get closer to the demand spike; an additional unit of capacity later in the congested period will save us from having to hold inventory in the first period of the congested interval and carry it until the current period.

In contrast to the LP model, however, the decision to hold FGI in the CF model is linked to the holding cost of WIP. The LP model will not hold FGI until capacity in a period is fully utilized since WIP costs are not modelled. In the CF model, however, the amount of WIP required to maintain high levels of utilization will eventually require very high levels of additional WIP for very little increase in output. In this situation, eventually the cost of additional WIP will exceed the cost of carrying FGI for several periods. Thus the CF model may hold FGI at utilization levels less than one.

We now examine the behavior of the system in non-congested periods.

### Table 1

<table>
<thead>
<tr>
<th>Utilization</th>
<th>( W_t )</th>
<th>( X_t )</th>
<th>( R_t )</th>
<th>( l_t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idle periods</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Non-congested periods</td>
<td>Low</td>
<td>&gt;0</td>
<td>&gt;0</td>
<td>&gt;0</td>
</tr>
<tr>
<td>Congested periods</td>
<td>High</td>
<td>&gt;0</td>
<td>&gt;0</td>
<td>≥0</td>
</tr>
</tbody>
</table>
of the horizon. In all cases we assume costs and the CF are constant over time.

We solve four different scenarios and compare the dual price of capacity obtained from each. All scenarios assume the same demand pattern and theoretical capacity of \( c_t = 15 \) units per period, but differ in the way capacity constraints are implemented. In Scenario 1 we make use of the classical LP formulation with a fixed capacity of 15 units as seen in Fig. 2. In the remaining three scenarios we use the CF formulation with different settings. Scenario 2 uses a CF with 4 segments, whereas Scenario 3 utilizes the same number of segments to represent a more congested resource (a CF that is more gradual). Scenario 4 uses the same CF as Scenario 3, but uses 8 linear segments, making our outer linearization a more exact representation of the underlying nonlinear concave function. Table 2 lists the CF parameters and Fig. 3 illustrates the three different CFs used.

![Fig. 2. Demand data for numerical examples.](image1)

![Fig. 3. CFs used in numerical examples.](image2)

<table>
<thead>
<tr>
<th>CF segment</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_c )</td>
<td>( \beta_c )</td>
<td>( x_c )</td>
<td>( \beta_c )</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>10</td>
<td>0.44</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
<td>12</td>
<td>0.11</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.18</td>
<td>4.9</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.08</td>
<td>7.65</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.02</td>
<td>11.02</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Parameters of linear functions approximating CFs.

We solved each scenario using the ILOG OPL (2007) environment. The dual prices obtained are plotted in Figs. 4 and 5. For the classical LP model, we report \( \sigma_t \) and for CF models we report \( \lambda_t \).

Since in the first half of the horizon the demand never reaches the maximum possible output in a period, the conventional LP model in Scenario 1 yields dual prices of zero for capacity in all periods in this portion of the horizon. In contrast, the dual prices yielded by all CF model Scenarios 1–3 are positive whenever production takes place, and increase or decrease based on resource utilization even though utilization is well below one. This suggests that at utilization levels below one, the conventional LP representation of capacity provides limited dual information.

In Fig. 4 we observe that during periods 27–39 and 40–47, when utilization is high, CF dual prices are consistently higher than those from the LP model when FGI is present. In these periods the capacity constraint of the LP model is tight, and in Scenario 2, we use the dual prices corresponding to the segment \( c = 2 \) being tight. It is interesting, however, that the dual prices from the two scenarios coincide quite closely. At first sight, this appears anomalous; at high utilization we would expect long queues to form, and given the ability of CF models to capture this effect, the dual price of capacity to be considerably higher in CF models. In fact, a careful investigation of Fig. 5 reveals that as we add more segments to the piecewise linearization (Scenario 3 versus Scenario 4) the CF model becomes more sensitive to the queuing behavior and dual prices start to climb. These observations suggest that Scenario 2 underestimates the dual price of capacity in regions of high utilization due to the piecewise linear approximation and limited number of segments.

Examining the output from the first and second scenarios in Fig. 6 we see that the LP model runs longer at full capacity, while the CF model elects to hold FGI and run below full capacity in periods 28–33 and 40–48. This latter behavior is frequently observed in real systems, where full utilization is never achieved but inventories are frequently present.

These observations suggest that in order to obtain more accurate dual prices, the piecewise linearization of the CF must be more detailed in the region corresponding to higher utilization levels. For intuition, we can consider expression (47) for the dual prices when a single linear segment is tight at optimality. Since the slope \( \sigma_{\star}(t) \) of the clearing function is in the denominator, small changes in this parameter, especially as it approaches zero with higher values of \( W_{t-1} + R_t \), will lead to large changes in the
dual price. Hence the use of a limited number of linear segments will lead to underestimation of the dual price of capacity at high utilization levels, despite the fact that previous studies have not shown this to be a major issue when CF models are used purely for production planning.

8. Managerial insights

Our results suggest that in congested periods, the dual price of capacity in the CF model depends on the shape of the CF and the holding costs of WIP and FGI in a tightly coupled manner. The shape of the CF in our model embodies the characteristics of the production resource under consideration, effectively determining how much WIP is required for a given output level. Clearly, the dual price of capacity increases as the resource utilization increases, since both $\beta$ and $1/\lambda$ increase with utilization. Moreover, the dual price of capacity increases linearly when FGI is held in congested periods. In non-congested periods, where resource utilization is low, the solution is degenerate, and the CF model does not yield any interesting insights on dual behavior.

Relative to the body of work summarized earlier that utilizes dual prices of capacity in decision-making, the main feature of our paper is to point out the deficiencies of traditional LP models in estimating the value of capacity. We further show that CF formulations provide richer dual information that is consistent with queuing models and industrial experience. Therefore, decision makers should be aware of the problems associated with certain production planning models that do not capture queuing effects. This is of particular interest in applications where dual prices of resources may be informative, such as estimation of setup costs, dispatching, and overhead allocation mentioned earlier.

9. Conclusions

We have presented a comparison of the dual prices for capacity yielded by a conventional LP model and a model based on clearing functions under some simplifying assumptions. The dual prices generated by the CF model have the desirable property that positive dual prices are obtained at utilization levels less than one, while the conventional model yields zero prices. On the other hand, the use of piecewise linear approximation to the clearing function can result in underestimation of true prices at high utilization levels. Our results also show that the CF model exhibits more realistic behavior than the classical LP model, holding FGI at utilization levels of less than one.

The results in this paper show that the CF model effectively addresses several of the conceptual difficulties associated with classical LP models of production facilities. Extensive experimental studies in multi-stage, multi-product environments, i.e. Asmundsson et al. (2006) and Missbauer (2002), have shown that when the CF is fitted appropriately, the CF models yield significantly more realistic results than classical LP models. In more complex production-inventory systems with multiple stages and multiple products, multiple interactions between WIP, demand, CF shape, and finished goods inventories between production stages render concise analytical expressions for dual prices hard to obtain, although numerical values of dual prices are straightforward. A systematic study of the relationships between prices obtained from our CF model and those obtained by other methods, such as the queuing-based approaches of Banker et al. (1986) and Morton and Singh (1988), is also an interesting direction for future work.

Acknowledgements

The research of Reha Uzsoy was partially supported by the National Science Foundation under Grant number DMI-0559136 and a research grant from the Intel Research Council. The research of Yahya Fathi was partially supported by the National Science Foundation under Grant number DMI-0321635.

References

Riaño, G., 2003. Transient Behavior of Stochastic Networks: Application to Production Planning with Load-Dependent Lead Times. School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.


