1995

Volume and Capacity Interaction in Facility Design

Buss, Arnold H.; Lawrence, Stephen R.; Kropp, Dean H.

Taylor & Francis

Buss, A.H., Lawrence, S.R., and Kropp, D.H., Volume and Capacity Interaction in Facility Design, IIE Transactions, Vol. 26, No. 4, 1994, pp. 36-49.
http://hdl.handle.net/10945/60792

VOLUME AND CAPACITY INTERACTION IN FACILITY DESIGN

ARNOLD H. BUSS
Operations Research Department, Naval Postgraduate School, Monterey, CA 93943-5000

STEPHEN R. LAWRENCE
College of Business Administration, University of Colorado, Boulder, CO 80309-0419

DEAN H. KROPP
John M. Olin School of Business, Washington University, St. Louis, MO 63130

This paper addresses the joint facilities design problem of determining both demand and capacity with stochastic demand arrivals and stochastic processing throughput. Using a simple *M/M/*1 queueing model of a profit maximizing firm, we link marketing and production decision variables by recognizing appropriate congestion costs, and show that coordinated decision-making provides results superior to making demand and capacity decisions sequentially. Sensitivity analysis indicates that the model is robust with respect to its assumptions and parameters. An example illustrates the approach and demonstrates the application of the model.

■ In this paper we address the joint problem of determining demand volume and production capacity under conditions of uncertainty. By introducing congestion costs, we link the demand volume decision (typically determined by the Marketing function) with the capacity design decision (typically made by the Manufacturing/Operations function). Analysis of the resulting model shows that the usual pattern of subordinating the production capacity decision to the demand volume decision is suboptimal when demand arrivals and processing times are uncertain, and further demonstrates that ignoring congestion costs can lead to potentially expensive errors when making volume and capacity decisions.

There is increasing interest in both the research literature and in practice in coordinating manufacturing and marketing functions. Much has been written about the problems of "functional silos" in which functional specialties make important decisions in isolation from other interested parties (Hayes, Wheelwright, and Clark [12]), and about the need for greater functional coordination (Shapiro [18]). Presumably, coordinating activities across functional boundaries will result in lower costs, faster response to customers, greater flexibility, and better utilization of resources (Crittenden [4]). However, much of this literature is anecdotal without theoretical or analytic underpinning. A principal objective of this paper is to provide such support for the facilities planning problem.

A second objective of the paper is to add support to the growing literature which argues that congestion effects must be recognized when addressing production problems. Much of the traditional capacity design literature assumes, either implicitly or explicitly, that capacity utilization rates of 100 percent are feasible and, indeed, desirable. However, recognition that workflow congestion creates negative impacts is increasing (Fry and Blackstone [8]). Manufacturing plants have typically used direct labor and machine utilization as the principle measure of production performance (Eloranta [5]). While direct labor charges are a diminishing fraction of total production costs in most manufacturing firms, the cost of materials is increasing relative to total production costs, thus shifting managerial attention from the control of direct labor expenditures to the control of inventories. Large work-in-process (WIP) inventories, once considered desirable to buffer production and provide high utilization, are currently regarded as costly impediments to effective manufacturing by obscuring quality problems, increasing lead times, reducing flexibility, and tying-up expensive capital. Since workflow congestion leads to increased WIP inventories, attention is shifting to relieving congestion and hence reducing WIP.

To capture the interaction between demand volume and production capacity in facilities design and to include the effects of congestion, we develop a simple profit-maximizing model of an *M/M/*1 production system under assumptions of declining returns to scale for both demand and capacity. With this model the mean arrival rate represents expected demand volume, the mean service rate represents production capacity, and the mean time-in-system captures workflow congestion. We deliberately choose a simple and parsimonious model for this analysis in order to capture the aggregate effects of demand volume and production capacity decisions in the spirit of Manne [15]. Consequently, we do not address other facilities design considerations such as plant layout, product mix, technology selection, and so forth. Other objectives such as return on investment (ROI) could also be considered. Although we briefly consider this, we will assume that we are dealing with a profit-maximizing firm.

The balance of the paper is organized as follows: In the next section we discuss the relevant capacity design literature, for which the decision variables have been either the demand rate or the service rate. The following section contains our basic model, which has both the demand and service rates as decision variables. We then analyze the model and give results about profitability and optimal utilization. Sensitivity analysis is undertaken, where the effects of errors in key model parameters are considered, and we illustrate an appli-

0740-817X/94/$3.00x.00

cation of our model and demonstrate the benefits of simultaneous determination of volume and demand. Issues involving the implementation of our model and offering directions for future research are discussed, and concluding remarks are made.

Literature

Hillier [13] was the first to consider the combination of capacity utilization, facilities design, and congestion costs. He derived a number of now-classic expressions for optimal service rates, number of servers, and customer arrival rates with linear capacity costs, but did not address the problem of jointly determining both service rates and arrival rates. Subsequent research has typically assumed either the demand rate or the service rate to be fixed, or has considered both to be deterministic.

There is a growing body of literature which investigates capacity design decisions given an exogenously imposed demand distribution. Freidenfelds [7] studies the impact of congestion on the timing of capacity expansion decisions. Yao and Kim [23] examine the loading of work stations and the assignment of servers to relieve congestion in an open queueing network. Karmarkar, Kekre and Kekre [14] use an *M/G/c* queueing model with linear capacity costs to examine capacity and equipment levels in a manufacturing cell with variable equipment levels, multiple shifts, overtime, and batching. Bitran and Tirupati [2] investigate capacity/inventory tradeoffs using decomposition techniques for the design of open queueing networks. Vander Veen and Jordon [20] examines the problem of designing a multi-machine production facility to meet fixed demand subject to nonlinear capacity costs, and Pourbabai [17] investigates the optimal utilization of a finite capacity integrated assembly system.

A second line of research assumes capacity to be fixed, and sets demand (usually through pricing mechanisms) to balance utilization with congestion costs. Banker, Datar and Kekre [1] investigate the impact of congestion on management accounting decisions. They use an *M/G/c* queueing model to show that ignoring congestion effects can seriously distort product costing analysis, thus leading to acceptance of work at unprofitable prices. In their model, plant capacity is considered to be fixed, while customer arrival rates are implicitly controlled through pricing mechanisms.

A final related stream of research examines the joint problem of determining production and demand levels, but typically assumes both to be deterministic and thus does not include congestion costs. Pekelman [16] investigates a model with dynamic inventory, price, and production decisions in which demand is a function of price and capacity is unlimited. Thompson, Sethi and Tang [19] and Feichtinger and Hartl [6] extend Pekelman's model by including nonlinear cost and demand functions. Gaimon [9] uses optimal control theory to determine price, capacity acquisition, production, mix, and inventory policies where demand is a function of price.

Other relevant work regarding the design of queueing systems is summarized in Crabill, Gross and Magazine [3]. Previous work has not considered the capacity design problem when both demand and production are stochastic — ours is apparently the first paper to investigate this joint problem.

*M/M/*1 Volume and Capacity Model

Our objective is to specify both the demand volume and the production capacity of a single-station production facility providing one or more products for some market. This production facility might be a single machine, work-center, or department, or could be a simple single-stage plant. Our analysis is not limited to manufacturing plants, but is equally applicable to service facilities such as retail stores, health care delivery offices, communications nodes.

Decision Variables. During the capacity design process we address, two decisions must be made: the capacity of the plant, and the volume of demand processed by the facility. We take as our decision variables both the fixed production *capacity* (defined as the mean processing rate μ) of the facility under consideration, and the demand *volume* (defined as the mean demand rate λ) of business that will be processed by the facility. Decision variables λ and μ are mean rates; realized demand and capacity during any period are assumed to be stochastic. The production capacity decision is impacted by choice of process, number of machines, plant layout, material handling methods, information technologies, and scope of manufacturing. The volume decision is typically considered a marketing decision and is affected by price, advertising level, promotion, sales effort, and distribution. Note that in most other operations management and industrial engineering literature, demand volume is taken as an exogenously specified constraint rather than as a decision variable.

Profit Maximizing Objective. The objective of our model is to maximize period profits Π defined as period contribution $\mathcal{M}$ less period capacity costs $\mathcal{K}$ and period congestion costs $\mathcal{F}$ so that

$$\Pi = \mathcal{M} - \mathcal{F} - \mathcal{K}. \quad (1)$$

Each of these terms is described in detail below.

First, however, note that the model requires that the cash-flow streams of all relevant revenues and costs be converted to equivalent *level* cash-flow streams (an annuity). This is accomplished using standard discounting techniques. For example, suppose that fixed costs of $10,000 are required to install some level of capacity μ. If the expected life of the capacity is indefinite, then the $10,000 can represent the present value of a perpetuity. (If the capacity has a finite life, then an annuity calculation would be appropriate.) For example, if the risk-adjusted opportunity cost of capital of the firm is 20% per annum, and assuming that the capacity will operate 2,000 hours per year, then the equivalent perpetuity is approximately $1 per working hour. This calculation is equivalent to that which an equipment leasing firm might undertake. If the ongoing cost of maintaining this capacity is $9 per hour (perhaps for fixed labor, maintenance, insurance, etc.), then the total period cost of installing and maintaining capacity level μ is $10 per hour. In a similar manner, all cash flows affected by the demand volume and production capacity decisions need to be converted to equivalent level cash flows and incorporated into the model. The choice of time period (minutes, hours, days, weeks) is arbitrary and immaterial so long as all parameters use the same units.

Contribution to Profit. We define contribution to profit as the difference between revenues and variable costs (exclusive of capacity costs) associated with a given level of expected demand. Components of contribution might include sales revenues, variable material and labor costs related to volume, and marketing and advertising expenditures (an example appears later). The contribution function $\mathscr{M}(\lambda)$ is assumed to be a pure function of the demand rate λ and is represented as a power function of demand: $\mathscr{M}(\lambda) = M\lambda^{\alpha}$, where M is a scale parameter, and α is the (constant) elasticity of contribution. This representation is commonly used in economics (an example is the familiar Cobb-Douglas function; Varian [21]), and allows increasing, constant, or decreasing returns to scale for demand by appropriate selection of values for α.

Capacity Costs. We define capacity costs to be the costs per time unit of providing an expected production capacity of μ. These *period* capacity costs include amortized fixed expenses such as equipment and building costs, plus other ongoing capacity-related expenses such as plant, maintenance, and fixed labor costs (an example appears in a later section). As with contribution, we represent capacity costs $\mathscr{K}(\mu)$ as a power function of capacity: $\mathscr{K}(\mu) = K\mu^{\beta}$, where K is a scale parameter, and β is the (constant) elasticity of capacity. This representation of capacity has been widely used in the capacity planning and capacity expansion literature (e.g., Manne [15] and Freidenfelds [7]).

Declining Returns to Scale. For both capacity costs and contribution we assume that there exist linear or declining returns to scale. For contribution function $\mathscr{M}$ this requires that $\alpha \leq 1$, and that $\beta \geq 1$ for the capacity cost function $\mathscr{K}$. We justify these declining return assumptions using a relevant range argument. In practice, both capacity costs and contribution functions will often follow an "S-shaped" curve. For such curves there are increasing returns to scale for low levels of volume and capacity, approximately linear returns for middle levels, and decreasing returns for high levels. Each of these regions can be individually represented by appropriate values of α and β in our model, although clearly our functions are not S-shaped.

In the case of increasing returns to scale, intuition suggests (and mathematics confirm) that the optimal policy would be to increase volume and capacity without limit (see Appendix). Similar results obtain when both volume and capacity have linear returns to scale. Only when the range of decreasing returns for demand and/or capacity is reached does a finite optimal solution exist. We will therefore restrict our analysis to this relevant range in which there are *decreasing* returns to scale in capacity or demand (or both), and for which a finite solution exists. Thus, we assume $\alpha \leq 1 \leq \beta$, with at least one strict inequality.

Congestion Costs. Congestion costs are the costs of maintaining work-in-process inventories and include warehousing and storage costs, materials handling expense, insurance costs, inventory tracking and expediting charges, capital opportunity costs, quality expenses arising from deteriorating in-process inventory, and other relevant inventory holding expenses. Consistent with most of the inventory management literature, we assume that congestion costs are a linear function of the time for which inventories are held (Hadley and Whitin [11]). Let F be the marginal cost of holding a production lot for one time period, and let $\mathscr{W}$ be the time an average production lot or job is held. Then if jobs are arriving at mean rate λ, the average period cost of congestion is $\mathscr{F} = F\lambda\mathscr{W} = F\mathscr{L}$, where $\mathscr{L} = \lambda\mathscr{W}$ is the average number of jobs in the system by Little's law.

We model congestion in the production facility using an *M/M/*1 queue with the usual assumptions of independent Poisson arrivals, exponential service times, first-come first-served discipline, and a single server. The mean steady-state number in system for an *M/M/*1 queue is $\mathscr{L}(\lambda,\mu) = \lambda/(\mu-\lambda)$ (Gross and Harris [10]), so that congestion function $\mathscr{F}$ becomes

$$\mathscr{F}(\lambda,\mu) = \frac{F\lambda}{\mu-\lambda}. \tag{2}$$

The *M/M/*1 queueing model is chosen for parsimony — it is the simplest model which captures the interaction of volume and capacity in determining congestion, and has the added benefit of analytic tractability. While few actual production systems satisfy all *M/M/*1 assumptions, sensitivity analysis in the example shown later demonstrate that our model is robust with respect to its assumptions.

*M/M/*1 *Volume and Capacity Model.* Our model is now fully specified and its objective function Π can be written:

$$\Pi(\lambda,\mu) = M\lambda^{\alpha} - \frac{F\lambda}{\mu-\lambda} - K\mu^{\beta}. \tag{3}$$

Implicit in the model is the constraint $\lambda < \mu$ indicating that mean steady-state demand cannot exceed mean steady-state capacity. It is further assumed that all finished goods can be sold immediately, that capacity is continuous and can be increased without limit, and that customer processing times vary with the inverse of capacity.

Several limitations of this model are worth noting. First, the model is for a single-stage production facility, and so is inappropriate for detailed analysis of stochastic manufacturing networks such as job-shops or flow-lines. However, it may be useful in these settings as an aid to determining the *aggregate* levels of production capacity and demand volume. We will discuss possible extensions of the model to queueing networks later. Second, implicit in the choice of an *M/M/*1 production system is the assumption that the coefficients of variation for both the demand process and the production process are unity, which precludes the direct modeling of variance effects. However, we show that the model is robust with respect to different coefficients of variation for both demand and production processes. Finally, the model assumes steady-state demand and production processes, and so does not directly accommodate dynamic changes in contribution function $\mathscr{M}$ (perhaps due to demand estimation errors, demand growth, or seasonality) or in capacity cost function $\mathscr{K}$ (perhaps due to learning or continuous-improvement effects). We later discuss how the model can be adapted for use in more dynamic environments.

Analysis of the Basic Model

In this section we analyze the basic model presented in the previous section. We examine first the special case without flow costs (i.e., no congestion), and subsequently the general model. Although we cannot obtain a closed-form solution for the general case, we will see that much can be said about the solution depending on the relative returns to scale.

Solution Without Congestion

If congestion costs are negligible (or are ignored), then $F=0$ and solution of the model is straightforward. Applying the usual constrained marginal analysis to this problem provides the solution $(\lambda^*_{NF}, \mu^*_{NF})$:

$$\lambda^*_{NF} = \mu^*_{NF} = \left(\frac{M\alpha}{K\beta}\right)^{1/(\beta-\alpha)}, \tag{4}$$

which is profit-maximizing providing $\beta > \alpha$. In the case that $\beta \leq \alpha$, the optimal decision would be to build an infinitely large plant serving infinite demand. Positive profits will be made in the range:

$$0 < \mu < \left(\frac{M}{K}\right)^{1/(\beta-\alpha)}, \tag{5}$$

where we assume that $\lambda = \mu$. It is clear that in the absence of congestion costs capacity utilization will always be one. Indeed, if any firm with the profit function in Equation (3) has $\mu > \lambda$, then profits can be increased by increasing λ until $\mu = \lambda$. That is, $\varrho \equiv \lambda/\mu = 1$. This situation is not observed, however; capacity utilization is almost always less than unity; in recent times aggregate capacity utilization in the industrial sector has been well below 85% (for instance, see Banker, Datar, and Kekre; [1]).

Solution With Congestion

If $F > 0$, then the solution obtained above is suboptimal, since the resulting congestion would make flow costs arbitrarily large. To guarantee a finite solution, we assume that $\alpha \leq 1 \leq \beta$, where at least one inequality is strict. The first-order conditions for an optimal solution are:

$$\Pi_\lambda = M\alpha\lambda^{\alpha-1} - \frac{F\mu}{(\mu-\lambda)^2} \equiv 0 \tag{6}$$

$$\Pi_\mu = \frac{F\lambda}{(\mu-\lambda)^2} - K\beta\mu^{\beta-1} \equiv 0. \tag{7}$$

We also have, from $\lambda\Pi_\lambda + \mu\Pi_\mu \equiv 0$, that $M\alpha\lambda^\alpha = K\beta\mu^\beta$. Thus, the optimal capacity μ^* is a solution to $\phi(\mu) = 0$, where

$$\phi(\mu) = \mu^{(\alpha\beta+\beta-\alpha)/2\alpha} - \left(\frac{M\alpha}{K\beta}\right)^{1/\alpha} \mu^{(\alpha\beta-\beta+\alpha)/2\alpha} + \left(\frac{F}{K\beta}\right)^{1/2} \left(\frac{M\alpha}{K\beta}\right)^{1/2\alpha} \tag{8}$$

(see Appendix). The behavior of ϕ depends on the sign of the exponent of μ in the second term of Equation (8) giving rise to three cases: $\alpha < \beta/(\beta+1)$, $\alpha = \beta/(\beta+1)$, and $\beta/(\beta+1) < \alpha$.

Case 1: $\alpha < \beta/(\beta+1)$ — *Guaranteed Solution.* In this case, diseconomies of scale are sufficient to insure a unique positive local maximum at a positive level. This solution is given by $(\lambda^*, \mu^*) = ((K\beta/M\alpha)^{1/\alpha}(\mu^*)^{\beta/\alpha}, \mu^*)$, where μ^* is the solution to $\phi(\mu) = 0$. Note that for this case it is possible to have $\beta < 1$, that is, *increasing* returns in volume, provided the decreasing returns in capacity are sufficient (see Appendix).

Case 2: $\alpha = \beta/(\beta+1)$ — *Possible Unique Solution.* The likelihood that this condition will *exactly* hold is, of course, rather small, but we include it for completeness. As in Case 1, $\phi'(\mu) > 0$, but $\phi(0) \leq 0$ or $\phi(0) > 0$, depending on the sign of

$$\left(\frac{M\alpha}{K\beta}\right)^{1/2\alpha} - \left(\frac{F}{K\beta}\right)^{1/2}. \tag{9}$$

If Expression (9) is greater than zero, then there is always a solution μ^* to $\phi(\mu) = 0$ which, as with the previous case, results in an optimal solution. On the other hand, if Expression (9) is less than or equal to zero, then there is no non-zero optimal solution; profits may always be increased by decreasing the scale of operations to levels that are arbitrarily close to zero (i.e., $\lambda \rightarrow 0$, $\mu \rightarrow 0$).

Observe that the condition involving Expression (9) for an optimal solution is equivalent to an upper bound on the unit flow costs F:

$$F < K\beta\left(\frac{M\alpha}{K\beta}\right)^{1/\alpha}. \qquad (10)$$

Condition (10) shows that congestion costs must be sufficiently small in order for a non-zero optimal solution to exist.

Case 3: $\alpha > \beta/(\beta+1)$ — *Possible Unique Solution.* In this case, a non-zero solution may or may not be optimal. The function ϕ is unimodal, having a unique minimum at

$$\bar{\mu} = \left(\frac{\alpha\beta-\beta+\alpha}{\alpha\beta+\beta-\alpha}\right)^{\alpha/(\beta-\alpha)}\left(\frac{M\alpha}{K\beta}\right)^{1/(\beta-\alpha)} \qquad (11)$$

(see Appendix). Furthermore, $\phi(\mu) = 0$ will have two solutions, provided

$$F < \frac{4(\beta-\alpha)^2 K\beta}{\alpha\beta+\beta-\alpha}\,\bar{\mu}^{(\alpha\beta-\beta+\alpha)/\alpha}. \qquad (12)$$

Observe that if (12) is an equality, then there is a single solution, and if the inequality in (12) is reversed, then there are no solutions. If the latter, profits may always be increased by decreasing the scale of operations, so that the optimal solution is arbitrarily close to zero. Since $\bar{\mu}$ only depends on α, β, K, and M, inequality (12) shows that flow costs must be sufficiently small (relative to capacity costs) for a positive optimal solution to exist. On the other hand, flow costs that are too high can result in the optimal solution being (0,0). In this situation it is best to not produce at all or, if other considerations force production, to produce at the smallest possible level. Similar results can be obtained by focusing on λ rather than μ, as above. For the remainder of the paper we will assume that there is a finite, non-zero optimal solution.

Positive Profits. For the decreasing returns model to exhibit finite optima ($\mu < \infty$ and $\lambda < \infty$), it is necessary to have positive profits for some combination (λ, μ). Otherwise, operating at any positive level will be dominated by not producing or selling at all, with the consequent zero profits. It is therefore of interest to obtain conditions under which profits will be positive. For Cases 1 and 2 the conditions for positive profits are no more stringent than the basic conditions for optimality. For Case 3, the profit-maximizing solution is the larger of the two solutions to $\phi(\mu) = 0$, and the condition for profits to be positive is (see Appendix)

$$F < \frac{M(\beta-\alpha)^2\mu^{\alpha+1}}{\alpha\beta(\alpha\beta-\beta+\alpha)^{1-\alpha}}. \qquad (13)$$

Now consider $\Pi(\lambda^*(\mu),\mu)$, in which $\lambda^*(\mu)$ is the solution to (6) for fixed μ. The optimal solution is $(\lambda^*(\mu^*),\mu)$, where μ^* is the larger solution to $\phi(\mu) = 0$. If $\Pi(\lambda^*(\mu^*), \mu^*) > 0$, then there is an interval (μ_L, μ_U) in which $\Pi(\lambda^*(\mu),\mu) > 0$. That is, there is an interval in which positive profits are attainable.

Optimal Capacity Utilization. Our model demonstrates that high capacity utilization ϱ is not necessarily a profitable goal. Using the relation $\lambda = \varrho\mu$, the optimal capacity utilization is given by

$$\varrho^* = \lambda^*/\mu^* = \left(\frac{K\beta}{M\alpha}\right)^{1/\alpha}(\mu^*)^{(\beta-\alpha)/\alpha}. \qquad (14)$$

The condition of positive profits may be used to develop a lower bound on the optimal capacity utilization (see Appendix):

$$\varrho^* > 1-\frac{\beta-\alpha}{\alpha\beta}. \qquad (15)$$

The lower bound in Equation (15) is consistent with the firm's desire to utilize capacity as fully as possible. This bound represents the minimum utilization required to have positive optimal profits. On the other hand, the value of Π is $-\infty$ for $\varrho = 1$, indicating the futility of attempting to achieve 100% utilization levels. Clearly utilization above a certain level will drive profits negative through high congestion costs. Thus, there will be a range for capacity utilization within which the firm can profitably operate.

Sensitivity Analysis

We now discuss the sensitivity of the optimal capacity, volume, and utilization results to deviations from optimality in the decision variables and to mis-specification of the model or its parameters. We consider, in turn, contribution (M), the cost parameters (K and F), and the "returns to scale" parameters (α and β). Finally, we examine the sensitivity of the underlying $M/M/1$ model to changes in the coefficients of variation in both arrival and service processes. Throughout the section, we will assume that the conditions for a finite, non-zero optimum are met.

Comparative Statics. Since the optimal solution (λ^*,μ^*) solves the first order conditions, we may obtain the derivatives $\partial\mu^*/\partial M$ and $\partial\lambda^*/\partial M$ by differentiating both sides of (6) and (7) with respect to M and solving (see Appendix). The derivatives of ϱ^* are obtained by the relationship $\varrho^* = \lambda^*/\mu^*$, and those with respect to K and F are similarly obtained. The exact expressions for these derivatives are of less interest than their respective signs, which are:

$$\frac{\partial\mu^*}{\partial M} > 0;\quad \frac{\partial\lambda^*}{\partial M} > 0;\quad \frac{\partial\varrho^*}{\partial M} > 0. \qquad (16)$$

$$\frac{\partial\mu^*}{\partial K} < 0;\quad \frac{\partial\lambda^*}{\partial K} < 0;\quad \frac{\partial\varrho^*}{\partial K} < 0. \qquad (17)$$

$$\frac{\partial\mu^*}{\partial F} < 0;\quad \frac{\partial\lambda^*}{\partial F} < 0;\quad \frac{\partial\varrho^*}{\partial F} < 0. \qquad (18)$$

In (16) the signs of the first two derivatives are intuitive, since increasing unit contribution leads to a higher level of activity, and hence a higher demand rate. The corresponding increase in capacity is in reaction to the increase in desired volume. Also, optimal capacity utilization ϱ^* increases with increasing contribution M. The intuition behind this result is that as contribution M becomes large relative to congestion costs F, greater capacity utilization will be sustainable. Thus, the response to increased contribution is an increase in both volume and capacity, with an increased utilization level. However, μ^* increases faster than λ^* in this case. Observe further that the previous analysis shows that ϱ^* decreases from unity as congestion costs increase from zero, which is consistent with these comparative-statics results.

Similarly, in (17) we see that increasing unit capacity costs decreases the level of both capacity and volume. Capacity utilization also decreases, since capacity costs are now higher relative to congestion costs. From (18), increases in F also result in both lower capacity and volume levels. Capacity utilization levels also decrease, as expected, since congestion is associated with higher utilization levels and increasing congestion effects creates pressure to reduce congestion.

Deviations from Optimality. We now examine the consequences of deviations from (λ^*, μ^*) to demonstrate the robustness of the *M/M/*1 model. For a given scenario (contribution, costs, and returns-to-scale parameters), the profit function is relatively flat near the optimum. Figure 1 shows the percentage deviations from optimal profits as a function of percent deviation from optimal decisions for volume (λ) and capacity (μ). The parameters in Figure 1 are from the example in the following section. For capacity, deviations on the high side are not as serious as on the low side. Clearly it is better for the firm to overestimate its capacity requirements than underestimate them. If too little capacity is built, then the flow costs become large as congestion increases. Thus, while Figure 1 demonstrates that the profit function is relatively flat near the optimum, it also shows the dramatic impact of flow costs when capacity requirements are not met. Often, capacity can only be obtained in discrete units rather than continuously (see the example in the following section) and the actual capacity decision must be rounded up or down. Figure 1 demonstrates that it is preferable to round up for capacity decisions. The opposite situation exists for the volume decision: if too little demand is generated, then profitability is not hurt as badly as with too much volume, with its increased congestion and high flow costs. When faced with discrete demand volumes, it is better to err on the low side (i.e., round down) rather than the high side. Note, however, that these results are both consistent with the fact that too little congestion is better than too much congestion. That is, for capacity utilization it is better to round down than up.

Mis-Specified Parameters. Next, we examine the impact of mis-specifying the parameters of the model. Since these values must typically be estimated, they will probably be different from their actual values, and consequently the robustness of the model to these deviations is important. Figure 2 shows the percentage deviation from optimal profit as a function of percentage deviation from true parameter values. The deviation from optimality is small for the cost and contribution parameters (K, F, and M); as evidenced by the flat curves in Figure 2. Profits are only slightly more sensitive to deviations in K and M than F. This relative insensitivity of the model to estimation errors in congestion costs F is fortunate in that F will typically be the most difficult parameter to estimate.

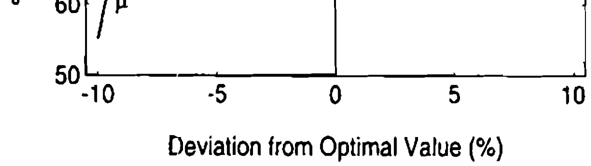

Figure 1. Sensitivity of profit to deviations in optimal volume and capacity. Example of degradation in profits when optimal volume λ or capacity μ is not achieved. For example, if mean volume is 10% greater than expected, profits decline by 30% (capacity assumed fixed). Here, M = 2005, K = 160, F = 250, α = 0.715, and β = 1.333.

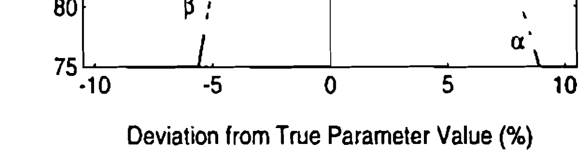

Figure 2. Sensitivity of profits to parameter mis-specification. Degradation in profits when "true" parameter values deviate from estimated values. For example, if true flowcosts F are 10% larger than estimated, profits will decline by about 0.5%. Here, M = 2005, K = 160, F = 250, α = 0.715, and β = 1.333.

The same figure shows the impact of mis-estimating the economy-of-scale parameters, α and β. The percent impact on profitability is much greater for these parameters. Furthermore, we see that it is better to underestimate α than to overestimate it, and that the reverse is true for β. Recall that increasing α gives rise to increasing economies of scale which will lead to increased capacity and profits. As Figure 2 indicates, the magnitude of this effect can be substantial.

Mis-Specified Model. Finally, we examine some consequences of incorrectly using the *M/M/*1 model for congestion. Since it is beyond the scope of this paper to examine *all* alternative models we will confine ourselves to the *GI/G/*1 queue, which can represent a broad range of single-station production facilities. This enables us to study the effect of changes in the coefficients of variation for processing (c_s) and inter-arrival times of demand (c_a). From the *GI/G/*1 heavy traffic approximation (Gross and Harris [10]), the mean steady-state number in a *GI/G/*1 system can be approximated as

$$\mathcal{L}^{(\mathcal{H})} \doteq \varrho + \frac{c_a^2 + c_s^2\varrho^2}{2(1-\varrho)}, \qquad (19)$$

for levels of high utilization (heavy traffic). While this expression strictly holds only in the limit as $\varrho \rightarrow 1$, it has been shown to be robust for utilization levels as low as $\varrho = 0.6$ in some contexts (Wein [22]). With Figure 3 we illustrate the sensitivity of the *M/M/*1 production model to changes in the coefficients of variation for arrival processes (c_a) and service processes (c_s). The *M/M/*1 model, of course, implicitly assumes that $c_a = c_s = 1$. As the figure shows, the *M/M/*1 model is relatively insensitive to processing-time cv for $c_s \leq 1.0$ and $c_s \approx 1.0$, the range in which we would expect many actual production processes to exist. But when both c_a and c_s are small (nearly deterministic) or both large (highly variable), the *M/M/*1 model differs significantly from the *GI/G/*1 model. In these situations, care must be taken when applying the results of the *M/M/*1 model.

Sequential vs Joint Decisions. As discussed earlier, the volume and capacity decisions are often made sequentially rather than jointly. We now consider the consequences of each decision being made sequentially starting from the zero flow cost solution of Equation (4). If there indeed exist flowcosts, this solution is infeasible and must be adjusted to attain feasibility. Two sequential approaches of attaining feasibility are to hold volume fixed and adjust capacity upward, or to hold capacity fixed and adjust volume downward. We term the first the MARKETING solution, since the capacity (production) decision is subordinated to the volume (marketing) decision, and the latter the PRODUCTION solution, since volume is subordinated to capacity. Figure 4 compares the sequential MARKETING and PRODUCTION solutions with the optimal (joint) solutions over a wide range of flowcosts F, using the parameters of the example in the following section. As flowcosts increase, the percentage differences in profits between the optimal (joint) solution and the sequential PRODUCTION and MARKETING decisions grow arbitrarily large. This particular result is interesting in light of the increasing emphasis on cycle time reduction and just-in-time production, indicating an increased need for joint capacity/volume decisions as flowcosts effectively increase. Similar figures can be obtained by varying capacity costs and contribution margins.

In summary, our model is robust with respect to deviations from the true basic cost and revenue parameters and is reasonably robust with respect to to demand and processing variances, but exhibits greater sensitivity to deviations from the true economy of scale factors. The implications are that economies of scale play an impor-

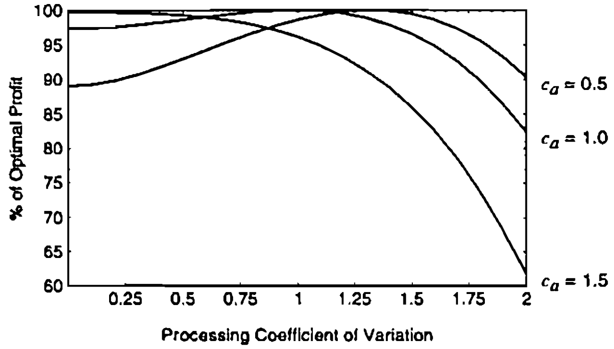

Figure 3. Sensitivity of profits to arrival and service processes. Degradation in profits when the coefficients of variation for arrival processes (c_a) or service processes (c_s) deviate from 1.0, the assumed coefficients of variation for the *M/M/*1 volume/capacity model. A heavy-traffic *GI/G/*1 approximation was used. An *M/G/*1 model produced a nearly identical curve with that for $c_s = 1.0$. Here, $M = 2005$, $K = 160$, $F = 250$, $\alpha = 0.715$, and $\beta = 1.333$.

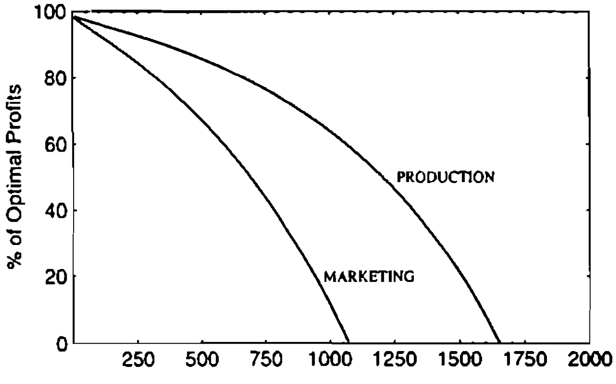

Figure 4. Optimal vs sequential decisions. Comparison of sequential MARKETING and PRODUCTION solutions as a percentage of optimal solutions obtained over a range of flowcosts F. Here, $M = 2005$, $K = 160$, $\alpha = 0.715$, and $\beta = 1.333$.

tant role in deciding the overall level of capacity and utilization, and that effort devoted to accurate measurement of these quantities is well spent. Finally, our model demonstrates the potential superiority of the joint over sequential approaches when making the capacity/volume decision.

An Example with Declining Returns to Scale

In this section we demonstrate with an example the benefit of simultaneous optimization of demand volume and production capacity compared with sequential optimization. The example is sufficiently complex so that several of the assumptions of the *M/M/*1 model are violated, but is also simple enough for a nearly exact solution to be found. This allows us to compare the quality of the approximate *M/M/*1 solution relative to the exact. Note that for many real-world problems an exact solution will be inaccessible, thereby necessitating the use of such approximations.

Consider a job-shop manufacturer that is expanding a profitable line of business. For the process in question, the manufacturer currently has a single machine ($k=1$) with a single-shift capacity of $\nu=50$ units per month, the costs of which are considered sunk. Additional machines can be acquired (purchased or leased) and operated for a cost equivalent to $1,500 per month (installation and maintenance included), and each requires one machine operator costing $1,000 per month to employ. The current facility has room for a total of four machines.

Expansion beyond $k=4$ machines would require the introduction of a second shift. Shift premiums and additional supervision would increase machine operator costs to $1,500 per month, while economies of scale (heat, power, maintenance) would reduce machine costs to $1,250 per month. From one to four machines could be used on the second shift, for a maximum shop capacity of eight machine-shifts. Further expansion would require the acquisition of an additional off-site building at a cost of $3,000 per month, would require a foreman at $2,000 per month, and ancillary services and supervision costing $3,000 per month. Note that capacity is not a continuous variable, as assumed by our model, but changes in discrete increments.

The contribution of additional business (revenues less variable material and marketing costs) is $125 for the first 50 units sold. Additional business commands a decreasing price and is increasingly expensive to secure. The next 100 units yield a contribution of $80 per unit; the following 100 units provide $65 per unit in contribution; sales between 250 and 400 would bring about $55 per unit; and sales beyond 400 units would yield a contribution of only $45 per unit. The costs of capacity and the contribution function are shown in Figure 5.

Jobs arrive to the facility as orders from customers having geometrically distributed order quantities with a mean of 10 units, with exponential times between orders. Let λ be the total number of units arriving per month, so that the mean number of orders per month will be $\lambda/10$. Orders are to be serviced on a first-come first-served basis. The processing time of each unit is effectively deterministic, so the time required to process each order has a geometric distribution with mean $10/\mu$, where $\mu = k\nu$ is monthly unit capacity. Order flow costs are $250 per month reflecting the importance of fast cycle times. The task of management is to determine the number of machines k to put in service, and the expected volume of business λ to solicit and accept, such that profits Π are maximized.

An Almost Exact Joint Solution. For this simple problem, a good numerical solution is possible using the actual contribution and capacity cost functions described above, together with congestion costs based on the expected number in system for an *M/M/k* multi-server queueing system (Gross and Harris [10]). This solution

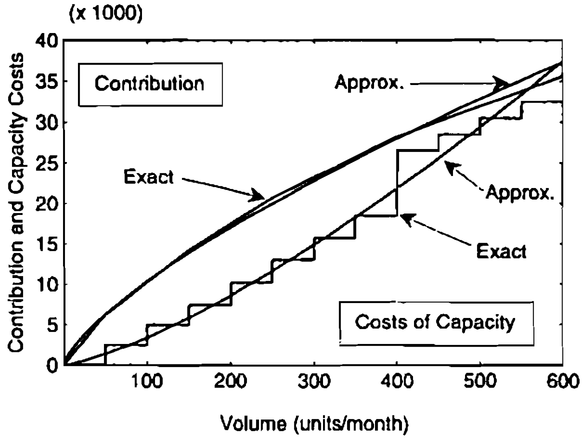

Figure 5. Example: Actual and approximate cost and revenue functions

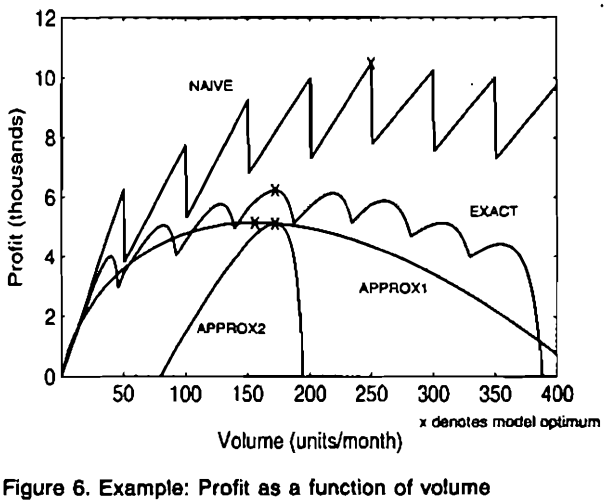

Figure 6. Example: Profit as a function of volume

is exact except that the geometric service distribution is approximated as exponential.

The resulting profit function can be maximized by solving for λ for each feasible value of k and choosing the best solution. Figure 6 illustrates profit as a function of volume λ, and shows that maximum profits of $\Pi = \$6{,}232$ are achieved with total capacity $\mu = 200$ ($k = 4$ machines), a monthly volume of $\lambda = 172$ units, with corresponding utilization of $\varrho = 0.86$. We label this the EXACT solution.

Approximate Joint Solution. We now apply the model previously developed in this paper to this example. First, the capacity cost function $\mathcal{K}(\mu)$ is approximated as the power function $7.430\mu^{1.333}$, the parameters of which are found via log-linear regression. Note that production is approximated as a continuously variable single-server rather than as multiple discrete-capacity servers. Similarly, the contribution function $\mathcal{M}(\lambda)$ is approximated by $386.147\lambda^{0.715}$. These approximate contribution and capacity cost functions are shown in Figure 5. Geometrically distributed batch processing times are approximated as exponential with mean $10/\mu$. Transforming $\mathcal{M}(\lambda)$ and $\mathcal{K}(\mu)$ to reflect batch arrivals, the capacity design problem facing the manufacturer is estimated as

$$\Pi(\hat{\lambda}, \hat{\mu}) = 2005\, \hat{\lambda}^{0.715} - 250\, \frac{\hat{\lambda}}{\hat{\mu} - \hat{\lambda}} - 160\, \hat{\mu}^{1.333} \quad (20)$$

where $\hat{\lambda} = \lambda/10$ is the mean batch interarrival rate and $\hat{\mu} = \mu/10$ is the mean batch processing rate.

Analysis and solution of this model proceeds as follows. First note that $\beta/(\beta+1) < \alpha < \beta$, which allows application of the results of Case 3. Expression (11) gives $\bar{\mu} = 3.667$, and inequality (10) is satisfied ($250 < 381.2$), verifying that a unique optimum solution to profit function in Expression (20) exists. Solving (20) numerically provides an optimal solution of $\lambda = 156$ and $\mu = 181$ units per month. The expected profits are $\Pi = \$5{,}160$ (APPROX1 solution), and optimal utilization of the facility will be $\varrho^* = 0.86$. Profit as a function of volume is plotted in Figure 6. Note that the APPROX1 solution is *infeasible*, since capacity must be acquired in discrete multiples of 50 units per month. Re-solving (20) with $\mu = 200$ yields a volume of $\lambda = 172$ (APPROX2 solution), expected profits of $\Pi = \$5{,}134$ (see Figure 6), and utilization of $\varrho = 0.86$.

Congestion Ignored. If congestion effects are ignored, expected profits are easily calculated as the difference between contribution and capacity costs, as shown earlier. Profit as a function of volume for this case is shown in Figure 6, which indicates that maximum profits are achieved with $\lambda = \mu = 250$ ($k = 5$) which provides *apparent* expected profits of $\Pi = \$10{,}375$ (NAIVE solution). Of course, actual profits will be far less than this.

Sequential Decisions. Volume and capacity of $\lambda = \mu = 250$ units per month could not be sustained in reality, since the resulting losses would be arbitrarily large due to congestion effects, requiring that either λ or μ be adjusted. First, suppose that $\lambda = 250$ is fixed (perhaps because of prior marketing commitments) so that production capacity must be adjusted to meet this demand level. With $\lambda = 250$, actual profit is maximized with $\mu = 300$, providing expected profits of $\Pi = \$5{,}766$ (MARKETING solution). In contrast, if production capacity is first fixed at $\mu = 250$ and λ optimized, volume is reduced to $\lambda = 218$ and expected profit (including congestion costs) is $\Pi = \$6{,}136$ (PRODUCTION solution).

Comparisons. Table 1 summarizes the results obtained for this example. Note that the table distinguishes between apparent profits (which are a function of the model used) and actual profits which would occur in reality (calculated using the exact model). In addition to profits, Table 1 includes figures for return-on-investment r defined as $r(\lambda, \mu) = M\lambda^{\alpha}/(F\lambda/(\mu - \lambda) + K\mu^{\beta}) - 1$, which illustrates the period investment required to achieve a given contribution level. Several patterns are apparent upon examining Table 1. First, the two approximate solutions APPROX1 and APPROX2 based on our *M/M/1* volume/capacity model provide decisions which are quite close to those of the exact *M/M/k* solution. Both provide machine utilizations of 86%, as does the EXACT solution, while APPROX2 yields the

Table 1. Comparison of Example Solutions

				Apparent		Actual	
Solution	λ	μ	ϱ	r	Π	r	Π
EXACT	172	200	0.86	0.66	\$6,232	0.66	\$6,232
APPROX1	156	181	0.86	0.56	5,160	Infeasible	
APPROX2	172	200	0.86	0.50	5,134	0.66	6,232
MARKETING	250	300	0.83			0.39	5,766
PRODUCTION	218	250	0.87			0.49	6,136
NAIVE	250	250	1.00	1.02	10,500	Infeasible	

Note: *Apparent* profits are the profits expected given the assumptions of the model used and its resulting volume/capacity decision. *Actual* profits are the profits that in fact would be realized if the model and its decision were implemented, and are calculated using the exact model.

optimal volume/capacity decision. Both approximations provide conservative estimates for profit and for return-on-investment.

A second pattern is that sequential volume/capacity decision processes, represented by the MARKETING and PRODUCTION solutions, provide results which are clearly inferior to the joint decision (EXACT). For the sequential MARKETING solution, profits are about 7.5% less than with the joint EXACT solution, optimal volume and capacity are overstated by 45% and 50% respectively, and the rate of return-on-investment is reduced by almost 41%. For the PRODUCTION solution, achievable profits are 2% less than with the EXACT solution, while optimal volume and capacity are overstated by 27% and 25% respectively, and return-on-investment is reduced by almost 26%. These results are of course predicated on our choice of model parameters — different values will result in quantitatively different but qualitatively similar results (see Figure 4, for example).

Finally, Table 1 shows that ignoring congestion effects (NAIVE) when making demand volume and production capacity decisions can be very costly. Compared with the "actual" EXACT solution, the "apparent," but *infeasible*, NAIVE solution overestimates optimal facility capacity by 25%, optimal volume of business by 45%, expected profits by 68%. Clearly the NAIVE solution would be impossible to implement and would, at a minimum, require disruptive adjustment to obtain a feasible solution were implementation attempted.

Implementation and Extensions

For many real problems, the *M/M/*1 model may be inadequate in representing the actual cost functions confronted, and a more general model of the form (1) is required. For example, flowcosts may be a non-linear function of time in system, or contribution and capacity cost functions may not be adequately represented by simple power functions. The production system may be far more complex than the assumptions of *M/M/*1 allow, requiring, for example, a queueing network model to adequately capture its detail. In this case, flow costs would be modeled using the expected throughput for the network. Analytic expressions are available for many examples of such networks, making such an extension viable. In these more complex cases, optimization of the resulting model may require either numerical solution or solution by simulation analysis. Even in such settings, the single server model we have considered here may be useful for determining the approximate aggregate level of capacity and demand. Furthermore, the principal results of our analysis are unchanged in these more complex situations: (1) joint optimization of demand volume and production capacity is preferred to sequential optimization, and (2) there exists an optimal utilization of capacity which is less than unity.

We turn now to implementation issues and possible extensions of the general model (1). First we consider changes in the structure of underlying demand. Demand volume is one of the decision variables of our general model, so that, in theory, demand can be maintained at desired mean steady-state level λ^* by appropriate manipulation of pricing, advertising, and sales policies. Demand of course will vary stochastically on a period-by-period basis, but its long-run mean is assumed to be constant. However, if the underlying structure of contribution function $\mathcal{M}$ changes due to competitive pressures, changes in the prices of factor inputs, or changes in consumer demand preferences, then demand volume λ^* may change over time and the model may eventually need to be reoptimized.

Given a revised contribution function $\mathcal{M}'$, how the model is reoptimized will depend on the capacity cost function $\mathcal{K}$. At one extreme, capacity may fixed and unchangeable so that μ is no longer a decision variable. Reoptimization then becomes the simpler problem of optimizing profits Π with respect to λ with fixed μ and $\mathcal{K} = 0$. At the other extreme, if capacity is completely flexible and transitory, then $\mathcal{K}$ remains unchanged from the original model and μ and λ are solved jointly, as before. In the intermediate case with some fixed and some flexible capacity, a modified cost function $\mathcal{K}'$ must be constructed to reflect the fixed portion of capacity. The model can then be reoptimized using $\mathcal{M}'$ and $\mathcal{K}'$. Similar analyses can be performed if there are changes in underlying cost functions $\mathcal{K}$ or $\mathcal{F}$.

Our results suggest several directions for further investigation. First, instead of a profit maximizing objective, a return on investment (ROI) objective could be used, since many firms use ROI measures to evaluate capacity expansion and other capital investment decisions. Second, the preceding discussion suggests how the volume/capacity decision can be adjusted over time to accommodate changing demand and capacity structures. The model does not directly incorporate uncertainty regarding mean steady-state demand or capacity, perhaps as might arise in the growth stage of a product-line's life cycle. One extension then would be to incorporate this additional uncertainty into the model, perhaps using a news vendor inventory type of analysis. A related extension would be to add seasonally variable demand to the model — the model would then be required to trade-off the costs and benefits of low utilization during slack seasons with those of the high utilizations during busy seasons. Another extension would be to extend our analysis to networks of queues in which multiple demand streams and multiple work-center capacities must be coordinated. Finally, our model assumes a strict upper bound on capacity represented by variable μ. In practice, capacity is often flexible and can be adjusted through the use of overtime and subcontracting, usually at additional expense. An interesting extension would be the incorporation of this flexibility into our model.

Summary

In this paper we have addressed the joint facilities design problem of simultaneously determining demand volume and production capacity under conditions of stochastic demand arrivals and processing throughput. Using a simple *M/M/*1 model of a production facility, we have demonstrated that coordinating demand volume and production capacity decisions provides results which are superior to sequential decisions. This provides analytic support for the current interest in functional coordination and team decision-making. Furthermore, the growing importance of WIP reduction and manufacturing cycle-time improvement demands that congestion costs be recognized when making demand volume and production capacity decisions. We have shown that these congestion costs link demand volume and production capacity decision variables, thus requiring their joint optimization. Failure to include relevant congestion effects significantly distorts the assessment of profitable volume/capacity options, obscures the need for coordination, and leads to potentially costly errors.

Acknowledgements

The authors would like to thank the Associate Editor and the referees for their constructive comments. The criticism of one referee was particularly insightful, and improved the paper considerably.

REFERENCES

[1] Banker, R. D., Datar, S. M., and Kekre, S., "Relevant Costs, Congestion and Stochasticity in Production Environments," *Journal of Accounting and Economics*, 10, 171-197 (1988).

[2] Bitran, G. R., and Tirupati, D., "Tradeoff Curves, Targeting and Balancing in Manufacturing Queueing Networks," *Operations Research*, 37(4), 547-564 (1989).

[3] Crabill, T. B., Gross, D., and Magazine, M. J., "A Classified Bibliography of Research on Optimal Design and Control of Queues," *Operations Research*, 25(2), 219-232 (1977).

[4] Crittenden, V. L., "Close the Marketing/Manufacturing Gap," *Sloan Management Review*, 41-52 (Spring 1992).

[5] Eloranta, E., "Managing Production for Customer Service and Plant Efficiency," *Industrial Engineering*, 20(3), 36-40 (1988).

[6] Feichtinger, G., and Hartl, R., "Optimal Pricing and Production in an Inventory Model," *European Journal of Operational Research*, 19, 45-56 (1985).

[7] Freidenfelds, J., *Capacity Expansion: Analysis of Simple Models with Applications*, Elsevier North Holland, New York (1981).

[8] Fry, T. D., and Blackstone, J. H., Jr., "Planning for Idle Time: A Rationale for Underutilization of Capacity," *International Journal of Production Research*, 26(12), 1853-1859 (1988).

[9] Gaimon, C., "Simultaneous and Dynamic Price, Production, Inventory and Capacity Decisions," *European Journal of Operational Research*, 35, 426-441 (1988).

[10] Gross, D., and Harris, C. M., *Fundamentals of Queueing Theory*, Second Edition, John Wiley and Sons, New York (1985).

[11] Hadley, J., and Whitin, T. M., *Analysis of Inventory Systems*, Prentice-Hall, Englewood Cliffs, NJ (1963).

[12] Hayes, R. H., Wheelwright, S. C., and Clark, K. B., *Dynamic Manufacturing: Creating the Learning Organization*, The Free Press, New York (1988).

[13] Hillier, F. S., "Economic Models for Industrial Waiting Line Problems," *Management Science*, 10(1), 119-130 (1963).

[14] Karmarkar, U., Kekre, S., and Kekre, S., "Capacity Analysis of a Manufacturing Cell," *Manufacturing Systems*, 6(3), 165-175 (1987).

[15] Manne, A. S., "Capacity Expansion and Probabilistic Growth," *Econometrica*, 29(4), 632-649 (1961).

[16] Pekelman, D., "Simultaneous Price-Production Decisions," *Operations Research*, 22, 788-794 (1974).

[17] Pourbabai, B., "Optimal Utilization of a Finite Capacity Integrated Assembly System," *International Journal of Production Research*, 28(2), 337-352 (1990).

[18] Shapiro, B. P., "Can Marketing and Manufacturing Coexist?" *Harvard Business Review* (1977).

[19] Thompson, G. L., Sethi, S. P., and Teng, J. T., "Strong Planning and Forecast Horizons for a Model with Simultaneous Price and Production Decisions," *European Journal of Operational Research*, 16, 378-388 (1984).

[20] Vander Veen, D. J., and Jordon, W. C., "Analyzing Tradeoffs Between Machine Investment and Utilization," *Management Science*, 35(10), 1215-1226 (1989).

[21] Varian, H. R., *Microeconomic Analysis*, Second Edition, W. W. Norton & Company, New York (1984).

[22] Wein, L. M., "Capacity Allocation in Generalized Jackson Networks," Working Paper #2018-88, MIT Sloan School of Management, Cambridge, MA (1988).

[23] Yao, D. D., and Kim, S. C., "Reducing Congestion in a Class of Job Shops," *Management Science*, 33(9), 1165-1172 (1987).

Appendix

Optimality Conditions

First, we determine conditions under which there is an optimal profit solution by analyzing properties of the function ϕ and a related function θ which pertain to the optimal μ and λ, respectively. The first-order condition for λ (6) may be written as:

$$(M\alpha)^{1/2}\lambda^{(\alpha-1)/2} = \frac{(F\mu)^{1/2}}{\mu-\lambda}$$

$$\lambda^{(\alpha+1)/2} - \mu\lambda^{(\alpha-1)/2} + \left(\frac{F\mu}{M\alpha}\right)^{1/2} = 0. \qquad (21)$$

Denote the left side of (21) by $\theta(\lambda)$. We now show that for $\alpha < 1$ there is a unique solution, $\lambda^*(\mu)$, to $\theta(\lambda) = 0$. Since $\alpha < 1$, we have $\theta(0) = -\infty$, and $\theta(\infty) = \infty$. Furthermore,

$$\theta'(\lambda) = \frac{\alpha+1}{2}\lambda^{(\alpha-1)/2} + \frac{1-\alpha}{2}\lambda^{(\alpha-3)/2} > 0. \quad (22)$$

Thus, since θ is strictly increasing, continuous, and has values of both signs, it must have a unique zero.

Next, consider the function ϕ (see Equation (8)). From (6), we have

$$\mu = \lambda + \left(\frac{F\mu}{M\alpha\lambda^{\alpha-1}}\right)^{1/2}. \qquad (23)$$

Furthermore, considering both first order conditions (6) and (7), we have $\lambda\Pi_\lambda + \mu\Pi_\mu \equiv 0$, since both $\Pi_\mu \equiv 0$ and $\Pi_\lambda \equiv 0$. Thus, $M\alpha\lambda^\alpha \equiv K\beta\mu^\beta$ and

$$\lambda = \left(\frac{K\beta}{M\alpha}\right)^{1/\alpha} \mu^{\beta/\alpha}. \tag{24}$$

Substituting (24) into (23),

$$\mu = \left(\frac{K\beta}{M\alpha}\right)^{1/\alpha} \mu^{\beta/\alpha} + \frac{F^{1/2}(K\beta)^{(1-\alpha)/2\alpha}}{(M\alpha)^{1/2\alpha}} \mu^{(\alpha-\alpha\beta+\beta)/2\alpha}. \tag{25}$$

Dividing through by $(K\beta/M\alpha)^{1/\alpha}\mu^{(\alpha-\alpha\beta+\beta)/2\alpha}$ and rearranging, we get

$$\mu^{(\alpha\beta+\beta-\alpha)/2\alpha} - \left(\frac{M\alpha}{K\beta}\right)^{1/\alpha} \mu^{(\alpha\beta-\beta+\alpha)/2\alpha} + \left(\frac{F}{K\beta}\right)^{1/2} \left(\frac{M\alpha}{K\beta}\right)^{1/2\alpha} = 0. \tag{26}$$

Since $\phi(\mu)$ is the left side of (26), the optimal capacity, μ^*, is a solution to $\phi(\mu) = 0$.

Next, consider $\Pi(\lambda^*(\mu),\mu)$, where $\lambda^*(\mu)$ is the solution to $\theta(\lambda) = 0$ for given μ. The optimal μ, if it exists, is a solution of $\Pi_\mu(\lambda^*(\mu),\mu) = 0$, since

$$\begin{aligned}\frac{\partial\Pi(\lambda^*(\mu),\mu)}{\partial\mu} &= \Pi_\lambda(\lambda^*(\mu),\mu)\frac{d\lambda^*(\mu)}{d\mu} + \Pi_\mu(\lambda^*(\mu),\mu)\\ &= \Pi_\mu(\lambda^*(\mu),\mu),\end{aligned}$$

since $\Pi_\lambda(\lambda^*(\mu),\mu) = 0$. Now we show that the sign of $\phi(\mu)$ is the opposite sign of $\Pi_\mu(\lambda^*(\mu),\mu)$. Since $\lambda^*(\mu)$ solves the first order condition for λ, we have

$$\frac{F\lambda^*(\mu)}{(\mu-\lambda^*(\mu))^2} = \frac{M\alpha\lambda^*(\mu)^\alpha}{\mu}. \tag{27}$$

Substituting (27) into $\Pi_\mu(\lambda^*(\mu),\mu)$, we have

$$\begin{aligned}\Pi_\mu(\lambda^*(\mu),\mu) &= \frac{M\alpha\lambda^\alpha}{\mu} - K\beta\mu^{\beta-1}\\ &= (M\alpha\lambda^*(\mu)^\alpha - K\beta\mu^\beta)/\mu. \end{aligned}\tag{28}$$

Now,

$$\theta\left(\left(\frac{K\beta}{M\alpha}\right)^{1/\alpha}\mu^{\beta/\alpha}\right) = \left(\frac{K\beta}{M\alpha}\right)^{(\alpha+1)/\alpha}\mu^{\beta(\alpha+1)/2\alpha}$$

$$-\left(\frac{K\beta}{M\alpha}\right)^{(\alpha-1)/\alpha}\mu^{(\alpha\beta-\beta+\alpha)/2\alpha} + \left(\frac{F\mu}{M\alpha}\right)^{1/2} = \left(\frac{K\beta}{M\alpha}\right)^{1/2\alpha}\phi(\mu).$$

Thus, if $\phi(\mu) < 0$, then $\lambda^*(\mu) > (K\beta/M\alpha)^{1/\alpha}$, and hence $M\alpha\lambda^*(\mu)^\alpha - K\beta\mu^\beta > 0$, in which case $\Pi_\mu(\lambda^*(\mu),\mu) > 0$ from (28). On the other hand, if $\phi(\mu) < 0$, then $\lambda^*(\mu) < (K\beta/M\alpha)^{1/\alpha}$, $M\alpha\lambda^*(\mu)^\alpha - K\beta\mu^\beta < 0$, and $\Pi_\mu(\lambda^*(\mu),\mu) < 0$.

We will now consider, in order, each of the three cases in the section Solution With Congestion.

Case 1: $\alpha < \beta/(\beta+1)$. In this case, $\alpha\beta-\beta+\alpha < 0$, so $\phi(0) = -\infty$. Since $\alpha\beta+\beta-\alpha > 0$, $\phi(\infty) = \infty$, so $\phi(\mu) = 0$ has at least one solution. Since

$$\phi'(\mu) = \frac{\alpha\beta+\beta-\alpha}{2\alpha}\mu^{(\alpha\beta+\beta-\alpha)/2\alpha-1} - \frac{\alpha\beta-\beta+\alpha}{2\alpha}\mu^{(\alpha\beta-\beta+\alpha)/2\alpha-1} > 0, \tag{29}$$

there is a unique solution μ^* to $\phi(\mu) = 0$. Also, since $\phi(\mu) < 0$ for $\mu < \mu^*$ and $\phi(\mu) > 0$ for $\mu > \mu^*$, $\Pi(\lambda^*(\mu),\mu)$ is increasing on $[0,\mu^*)$ and decreasing on (μ^*, ∞). Thus, $(\lambda^*(\mu^*),\mu^*)$ is the global maximum of $\Pi(\lambda,\mu)$. From $M\alpha(\lambda^*)^\alpha = K\beta(\mu^*)^\beta$, the optimal solution may also be written $((K\beta/M\alpha)^{1/\alpha}(\mu^*)^{\beta/\alpha},\mu^*)$.

Case 2: $\alpha = \beta/(\beta+1)$. Here, we have

$$\phi(\mu) = \mu^\beta - \left\{\left(\frac{M\alpha}{K\beta}\right)^{1/\alpha} - \left(\frac{F}{K\beta}\right)^{1/2}\left(\frac{M\alpha}{K\beta}\right)^{1/2\alpha}\right\}. \tag{30}$$

Thus, if

$$\left(\frac{M\alpha}{K\beta}\right)^{1/2\alpha} \leq \left(\frac{F}{K\beta}\right)^{1/2} \tag{31}$$

then there is no positive solution to $\phi(\mu) = 0$, whereas if

$$\left(\frac{M\alpha}{K\beta}\right)^{1/2\alpha} > \left(\frac{F}{K\beta}\right)^{1/2} \tag{32}$$

then there is a unique solution μ^* given by

$$\mu^* = \left(\frac{M\alpha}{K\beta}\right)^{1/2\alpha\beta}\left\{\left(\frac{M\alpha}{K\beta}\right)^{1/2\alpha} - \left(\frac{F}{K\beta}\right)^{1/2}\right\}^{1/\beta}. \tag{33}$$

Since in this case $\phi(0) < 0$ and ϕ is strictly increasing, $\Pi(\lambda^*(\mu),\mu)$ is increasing for $\mu < \mu^*$ and decreasing for $\mu > \mu^*$. Thus, $(\lambda^*(\mu),\mu)$ is a global maximum of $\Pi(\lambda,\mu)$.

Case 3: $\alpha > \beta/(\beta+1)$. For this case, both $\alpha\beta+\beta-\alpha$ and $\alpha\beta-\beta+\alpha$ are positive. Hence, $\phi(0) > 0$ and $\phi(\infty) > 0$. Furthermore, on $(0,\infty)$ there is a unique minimum $\tilde{\mu}$ of ϕ, which can be seen by noting that $\phi''(\tilde{\mu}) < 0$. From the expression for ϕ' in (29), we can solve for $\tilde{\mu}$ to obtain (11).

Now, if $\phi(\tilde{\mu}) \geq 0$, then there is no positive optimal solution, since $\Pi(\lambda^*(\mu),\mu)$ is decreasing on $(0,\infty)$(except at $\tilde{\mu}$ in the case of equality). On the other hand, if $\phi(\tilde{\mu}) < 0$, then $(\lambda^*(\mu^*),\mu^*)$ is a local maximum of Π, where μ^* is the *larger* of the two solutions of $\phi(\mu) = 0$. If in addition $\Pi(\lambda^*(\mu^*),\mu^*) > 0$, then $(\lambda^*(\mu^*),\mu^*)$ is the *global* maximum of Π. To see this, denote the smaller solution to $\phi(\mu) = 0$ by μ_S^*, and note that $\phi(\mu) < 0$ on the interval (μ_S^*,μ^*) and $\phi(\mu) > 0$ on the intervals $(0,\mu_S^*)$ and (μ^*, ∞). Thus, $\Pi(\lambda^*(\mu),\mu)$ is decreasing on $(0,\mu_S^*)$, increasing on (μ_S^*,μ^*), and decreasing again on (μ^*, ∞), and μ^* is a local maximum. Since $\Pi(0,0) = 0$, the point (0,0) is a maximum of $\Pi(\lambda^*(\mu),\mu)$ on $(0,\mu_S^*)$. Thus, if $\Pi(\lambda^*(\mu^*),\mu^*) > 0$, then $(\lambda^*(\mu^*),\mu^*)$ is the global optimum.

Positive Profits. We see that the condition for which profits are positive are important in determining the existence of an optimal positive solution. We will now examine some conditions on the parameters for which this will hold and, since flow costs are probably the most difficult to determine accurately, focus on conditions on F that will ensure a positive solution.

As shown above, in Case 1 there will always be a positive solution. In Case 2 there will be a (positive profit) solution providing (32) holds. Equivalently, for Case 2 there will be a positive solution iff (10) holds.

For Case 3 the situation is a bit more complicated due to the two conditions required. First, μ^* is a *local* optimum iff $\phi(\tilde{\mu}) < 0$. Substituting the expression for $\tilde{\mu}$ in Equation (11) into this inequality and simplifying, we have a local optimum iff (12) holds.

Finally, we determine whether profits are positive for (λ^*, μ^*). From the first order conditions for λ and μ:

$$M\lambda^{\alpha} = \frac{F\lambda\mu}{\alpha(\mu-\lambda)^2} \tag{34}$$

$$K\mu^{\beta} = \frac{F\lambda\mu}{\beta(\mu-\lambda)^2}, \tag{35}$$

so

$$\Pi(\lambda^*, \mu^*) = \frac{F\lambda\mu}{\alpha(\mu-\lambda)^2} - \frac{F\lambda}{(\mu-\lambda)} - \frac{F\lambda\mu}{\beta(\mu-\lambda)^2}$$
$$= \frac{F\lambda}{\alpha\beta(\mu-\lambda)^2}(\alpha\beta\lambda - (\alpha\beta-\beta+\alpha)\mu).$$

Thus, $\Pi(\lambda^*, \mu^*) > 0$ iff $\lambda^* > \mu^*(\alpha\beta-\beta+\alpha)/\alpha\beta$. Now,

$$\theta\left(\frac{\alpha\beta-\beta+\alpha}{\alpha\beta}\mu\right) = -\frac{\beta-\alpha}{\alpha\beta}\left(\frac{\alpha\beta-\beta+\alpha}{\alpha\beta}\right)^{(\alpha-1)/2}\mu^{(\alpha-1)/2}$$
$$+\left(\frac{F\mu}{M\alpha}\right)^{1/2}$$

which is negative iff

$$F < \frac{M(\beta-\alpha)^2\mu^{\alpha+1}}{\alpha\beta(\alpha\beta-\beta+\alpha)^{1-\alpha}}. \tag{36}$$

Thus, if (36) holds for μ^*, then $\theta(\mu^*(\alpha\beta-\beta+\alpha)/\alpha\beta) < 0$, which in turn implies $\lambda^* > \mu^*(\alpha\beta-\beta+\alpha)/\alpha\beta$. From this, (15) immediately follows.

Comparative Statics

We turn to the comparative statics results in (16)-(18), assuming throughout that the conditions for a finite positive optimal solution are satisfied. The second order conditions are (dropping the $*$'s for ease of exposition):

$$\Pi_{\lambda\lambda}(\lambda,\mu) = -M\alpha(1-\alpha)\lambda^{\alpha-2} - \frac{2F\mu}{(\mu-\lambda)^3} < 0 \tag{37}$$

$$\Pi_{\mu\mu}(\lambda,\mu) = -\frac{2F\lambda}{(\mu-\lambda)^3} - K\beta(\beta-1)\mu^{\beta-2} < 0 \tag{38}$$

$$\Pi_{\lambda\mu}(\lambda,\mu) = \frac{F(\lambda+\mu)}{(\mu-\lambda)^3} > 0. \tag{39}$$

The fact that (λ^*, μ^*) is a maximum implies that

$$\Pi_{\lambda\lambda}\Pi_{\mu\mu} - \Pi^2_{\lambda\mu} > 0. \tag{40}$$

From the first order conditions at (λ^*, μ^*), we have, differentiating both sides of (6) and (7),

$$\Pi_{\lambda\lambda}\frac{\partial\lambda^*}{\partial M} + \Pi_{\lambda\mu}\frac{\partial\mu^*}{\partial M} + \frac{\partial\Pi_{\lambda}}{\partial M} = 0 \tag{41}$$

$$\Pi_{\lambda\mu}\frac{\partial\lambda^*}{\partial M} + \Pi_{\mu\mu}\frac{\partial\mu^*}{\partial M} + \frac{\partial\Pi_{\mu}}{\partial M} = 0. \tag{42}$$

Solving (41) and (42) simultaneously, using the fact that $\partial\Pi_{\lambda}/\partial M = \alpha\lambda^{\alpha-1}$ and $\partial\Pi_{\mu}/\partial M = 0$, we have, using (37)-(40) for the signs:

$$\frac{\partial\lambda^*}{\partial M} = \frac{-\alpha\lambda^{\alpha-1}\Pi_{\mu\mu}}{\Pi_{\lambda\lambda}\Pi_{\mu\mu} - \Pi^2_{\lambda\mu}} > 0 \tag{43}$$

$$\frac{\partial\mu^*}{\partial M} = \frac{\alpha\lambda^{\alpha-1}\Pi_{\lambda\mu}}{\Pi_{\lambda\lambda}\Pi_{\mu\mu} - \Pi^2_{\lambda\mu}} > 0. \tag{44}$$

Similarly, we have, using $\partial\Pi_{\lambda}/\partial K = 0$ and $\partial\Pi_{\mu}/\partial K = -\beta\mu^{\beta-1}$:

$$\frac{\partial\lambda^*}{\partial K} = \frac{-\beta\mu^{\beta}\Pi_{\lambda\mu}}{\Pi_{\lambda\lambda}\Pi_{\mu\mu} - \Pi^2_{\lambda\mu}} < 0 \tag{45}$$

$$\frac{\partial\mu^*}{\partial K} = \frac{\beta\mu^{\beta}\Pi_{\lambda\lambda}}{\Pi_{\lambda\lambda}\Pi_{\mu\mu} - \Pi^2_{\lambda\mu}} < 0. \tag{46}$$

Finally, for F we have

$$\frac{\partial\lambda^*}{\partial F} = \frac{(\mu\Pi_{\mu\mu} + \lambda\Pi_{\lambda\mu})/(\mu^*)^2}{\Pi_{\lambda\lambda}\Pi_{\mu\mu} - \Pi^2_{\lambda\mu}} \tag{47}$$

$$\frac{\partial\mu^*}{\partial F} = \frac{(\mu\Pi_{\mu\mu} + \lambda\Pi_{\lambda\mu})/(\mu^*)^2}{\Pi_{\lambda\lambda}\Pi_{\mu\mu} - \Pi^2_{\lambda\mu}}. \tag{48}$$

Thus,

$$(\mu-\lambda)^2\frac{\partial\lambda^*}{\partial F} = \mu\Pi_{\mu\mu} + \lambda\Pi_{\lambda\mu}$$
$$= -\frac{2F\lambda\mu}{(\mu-\lambda)^3} - K\beta(\beta-1)\mu^{\beta-1} + \frac{F\lambda(\lambda+\mu)}{(\mu-\lambda)^3}$$
$$= -K\beta^2\mu^{\beta-1} < 0.$$

Similarly,

$$(\mu-\lambda)^2\frac{\partial\mu^*}{\partial F} = -\mu\Pi_{\lambda\mu} - \lambda\Pi_{\lambda\lambda}$$
$$= -\frac{F\mu(\lambda+\mu)}{(\mu-\lambda)^3} - M\alpha(1-\alpha)\lambda^{\alpha-1} - \frac{2F\lambda\mu}{(\mu-\lambda)^3}$$
$$= M\alpha^2\lambda^{\alpha-1} > 0.$$

To determine the sensitivity of the optimal capacity utilization ϱ^*, we use the fact that $\varrho^* = \lambda^*/\mu^*$ and

$$\frac{\partial\varrho^*}{\partial M} = \left(\mu\frac{\partial\lambda^*}{\partial M} - \lambda\frac{\partial\mu^*}{\partial M}\right)/\mu^2, \tag{49}$$

with similar expressions for K and F. Substituting (43) and (44), into (49) we have

$$\frac{\partial\varrho^*}{\partial M} = -\frac{\alpha\lambda^{\alpha-1}(\mu\Pi_{\mu\mu} + \lambda\Pi_{\lambda\mu})}{\mu^2(\Pi_{\lambda\lambda}\Pi_{\mu\mu} - \Pi^2_{\lambda\mu})} > 0.$$

In a similar manner,

$$\frac{\partial \varrho^*}{\partial K} = -\frac{\beta\mu^{\beta-1}\,(\mu\Pi_{\lambda\mu} + \lambda\Pi_{\lambda\lambda})}{\mu^2(\Pi_{\lambda\lambda}\Pi_{\mu\mu} - \Pi_{\lambda\mu}^2)} < 0.$$

Finally, since $\partial\lambda^*/\partial F < 0$ and $\partial\mu^*/\partial F > 0$

$$\frac{\partial \varrho^*}{\partial F} = \left(\mu\frac{\partial\lambda^*}{\partial F} - \lambda\frac{\partial\mu^*}{\partial F}\right)/\mu^2 < 0.$$

Returns to Scale

We will demonstrate the impact of returns to scale on the results. Our discussion will not be comprehensive, but will give an indication of the role played by returns to scale in either volume or capacity.

First, consider the case of linear returns to scale for both volume ($\alpha = 1$) and capacity ($\beta = 1$). From the second order conditions (37) and (38), we see that $\Pi_{\lambda\lambda} < 0$ and $\Pi_{\mu\mu} < 0$, so Π is marginally concave in λ and in μ. However, from (39), we also have

$$\Pi_{\lambda\lambda}\Pi_{\mu\mu} - \Pi_{\lambda\mu}^2 = \frac{4F^2\lambda\mu}{(\mu-\lambda)^6} - \frac{F^2(\lambda+\mu)^2}{(\mu-\lambda)^6} = -\frac{F^2}{(\mu-\lambda)^4} < 0.$$

Thus, the solution to the first order conditions, while unique in this case, is in fact a saddle point. Indeed, as is intuitively clear, with linear returns, the scale of operations could be increased indefinitely with every-increasing profits. Clearly the situation remains the same if there are increasing returns for both volume and capacity. Now, it is clear that there will *eventually* be decreasing returns to scale in any real process due to situations associated with extremely large-scale operations: saturating markets, large unwieldy production systems, etc. Consequently, it is our implicit assumption that the firm has exploited whatever economies of scale that exist at low levels of operation and have reached the point of decreasing returns in one or both of volume and capacity. While we have dealt with the case in which there are declining returns in *both* volume and capacity, it is enough to have reached declining returns in only one of the two.

To analyze the cases in which there are linear or increasing returns in one and decreasing returns in the other would require a case by case analysis for each situation. We will not delve into all the various possibilities, but briefly sketch two. For Case 3 above, all the conditions for a finite positive optimal solution hold as long as $\beta/(\beta+1) < \alpha < 1$, regardless of the (positive) value of β. Thus, even if there are increasing returns for capacity, decreasing returns in volume may be sufficient for a finite optimum. Likewise, for Case 1 the results hold so long as the basic condition $\alpha < \beta/(\beta+1)$ is satisfied, regardless of the value of β.

Arnold H. Buss received his Ph.D. in Operations Research at Cornell University. His research interests include capacity planning, simulation, and the Operations/Marketing interface. He is a member of the Institute of Industrial Engineers.

Stephen R. Lawrence received his B.S. and M.S. degrees in Engineering Science from Purdue University, and M.S. and Ph.D. degrees in Operations Management from Carnegie-Mellon University. His current research interests include economic scheduling, capacity design, and international operations. He has taught operations strategy, and planning and control courses in MBA and Executive programs at The University of Colorado, Washington University, Carnegie Mellon University and the University of Pittsburgh. Previously, he worked in industry for eight years in operations management, quality management, and sales engineering.

Dean H. Kropp is the Dan Broida Professor of Operations and Manufacturing Management at the John M. Olin School of Business, Washington University, St. Louis, Missouri. His teaching, research, and consulting activities are in the areas of production planning and inventory control, plant location, quality management, and manufacturing strategy. His articles have been published in numerous journals, including *Management Science*, *Decision Sciences*, *IIE Transactions*, and *Journal of Operations Management*. Much of his research has examined ways to improve the effectiveness of production scheduling in a rapidly changing environment.