MPE — Volume 1, pp. 341-363
Reprints available directly from the publisher
Photocopying permitted by license only

ECONOMIC ANALYSIS OF PRODUCTION BOTTLENECKS

STEPHEN R. LAWRENCE

College of Business and Administration, University of Colorado, Boulder, CO 80309-0419

ARNOLD H. BUSS

Operations Research Department, Naval Postgraduate School, Monterey, CA 93943-5000

(Received 7 February 1995)

The management of bottlenecks has become a central topic in the planning and control of production systems. In this paper, we critically analyze bottlenecks from an economic perspective. Using a queueing network model, we demonstrate that bottlenecks are inevitable when there are differences in job arrival rates, processing rates, or costs of productive resources. These differences naturally lead to the creation of bottlenecks both for facilities design and demand planning problems. To evaluate bottlenecks from an economic perspective, we develop the notion of an "economic bottleneck," which defines resources as bottlenecks based on economic, rather than physical, characteristics.

AMS Nos.: 90B30, 90B50, 90B22
Keywords: Manufacturing, bottlenecks, capacity design, demand management, economic analysis

1. INTRODUCTION

In this paper we critically analyze production bottlenecks from an economic perspective, addressing important facilities-design and demand-planning problems. Using a queueing network model, we demonstrate that production bottlenecks are inevitable when there are differences in job arrival rates, processing rates, or costs of productive resources. We further analyze the impact of capacity and demand decisions on the location and characteristics of bottlenecks.

The results of this analysis have a number of important managerial implications. We demonstrate that bottlenecks necessarily arise when costs (profits) are minimized (maximized) and that attempts to control bottlenecks by balancing production will usually be economically counterproductive. We introduce the notion of an "economic bottleneck," which defines resources as bottlenecks based on economic, rather than physical, characteristics. This definition provides the basis for making recommendations for managing and relieving bottlenecks in order to improve the economic performance of productive capacity.

Thus we address the capacity allocation problem from a different perspective than that of most work previously done. Although our Jackson network model of the shop-floor dynamics is simpler than some others, we are able to obtain closed-form expressions,

rather than simply to present an algorithm. This allows us to obtain a much clearer picture of the *nature* of the solution and thus provide more insight. In particular, the economic bottleneck concept allows easy, managerially useful prescriptions for allocation of additional resources.

A survey of the research and pedagogical literature finds that there exists no clear consensus as to the definition of a "bottleneck" resource. Several of these definitions are:

- *Congestion points, or bottlenecks, primarily occur when manufacturing resources required in a given time period are unavailable.* [30]
- *A bottleneck is defined as any resource whose capacity is less than the demand placed upon it. A bottleneck, in other words, is a process that limits throughput.* [8]
- *Production bottlenecks are generally considered to be temporary blockades to increased output...(where) inventories build up at different places and different times.* [31]
- *A facility, function, department, etc., that impedes production...* [12]
- *A bottleneck operation ... is any operation that limits output.* [21]

Umble and Srikanth [33] make a distinction between "bottlenecks" and "constrained capacity resources," where the former is defined as "any resource whose capacity is equal to or less than the demand placed upon it" and the latter as "any resource which ... is likely to cause the actual flow of product through the plant to deviate from the planned product flow." Goldratt and Cox [13] and Schonberger and Knod [32] make similar distinctions.

Summarizing, there appear to be three principal definitions for bottleneck resources currently in use: A bottleneck resource is one for which (1) short-term demand exceeds capacity; (2) work-in-process (WIP) inventory is maximum; or (3) production capacity is minimum, relative to demand (*i.e.*, capacity utilization is maximal). We assign each definition its own name: the short-term definition, the inventory definition, and the production definition, respectively.

Short-Term Definition. In the long run, demand cannot exceed capacity—either work will increase without bound or there will be sufficient loss of business to reduce the demand rate below capacity. But demand can, and often does, exceed capacity in the short run. When these *short-term* bottlenecks occur, shop-floor control techniques (such as finite scheduling, priority sequencing, overtime, subcontracting, and so forth) must be exercised to alleviate the bottleneck.

Inventory Definition. In practice, bottlenecks are often identified by simply walking out on a shop floor and observing where the most work is waiting in queue for processing. The second definition of a bottleneck thus focuses on relative levels of work-in-process (WIP) inventories—that resource with the largest current WIP is defined to be the bottleneck. In a stochastic environment, the workcenter with the most WIP changes from time to time, and so the bottleneck is observed to randomly shift from workcenter to workcenter—a phenomenon which bedevils many managers (for an analysis of shifting bottlenecks, see [23]). Since work-in-process inventories are related to the busy period of the workcenter, they shift more slowly than do instantaneous arrival rates. Inventory bottlenecks are therefore more long-lived than are short-term bottlenecks, making them relevant over a middle-term time horizon.

Production Definition. Finally, for long-range planning, bottlenecks are those resources which, on average, are the greatest impediment to increased output or throughput. For long time horizons, capacity utilization is the most useful measure with which to identify these long-run bottlenecks, since those resources which are most highly utilized are those which over time most severely limit output and slow production.

These three definitions can be operationalized as follows. Let γ be the arrival rate of jobs to a workcenter or machine, and let μ be its capacity, expressed as a processing rate. Then a bottleneck exists when: (1) demand temporarily exceeds capacity (short-term definition); or (2) the number of jobs waiting in queue $L = L(\gamma,\mu)$ is maximum (inventory definition); or (3) the long-run utilization $\rho = \gamma/\mu$ is maximum (production definition). An alternative to utilization is long-run capacity cushion defined as $\mu - \gamma$. We use capacity cushion and utilization interchangeably as the context warrants. Note that these definitions are not mutually exclusive and that a particular workcenter may satisfy one or more of them at any given time. Since none of these definitions considers costs, revenues, or profitability of the firm, but focuses solely on the output of the process, we will refer to them as *production* bottlenecks. In contrast, when cost and revenues are taken into account, we can identify those resources which limit profitability. We will call such resources *economic* bottlenecks. As with production bottlenecks, we can identify short-, medium-, and long-term economic bottlenecks in a manner similar to the three types of production bottlenecks. Since the nature of the decisions we consider in this paper are long-term, we will use the long-term bottleneck definitions for both the production and economic bottlenecks. In this paper we demonstrate that the economic bottleneck and the production bottleneck do not necessarily coincide.

The remainder of this paper is organized as follows. The next section surveys the literature, and section 3 introduces the general profit maximization model for determining capacities or demand rates. Section 4 investigates optimal bottlenecks in facilities design where the design variable is production capacity. In section 5 we turn to the demand planning problem of determining optimal demand volumes when resource capacities are fixed. We introduce the notion of an *economic bottleneck* in section 6 and discuss its implications in managing bottlenecks. Section 7 presents a numerical example of our results, and concluding remarks are made in section 8.

2. LITERATURE REVIEW

The importance of understanding and managing production bottlenecks has been the focus of the proponents of the "OPT" philosophy (for example, [13]), more recently called the Theory of Constraints [14] or Synchronous Manufacturing [33]. This literature argues that bottlenecks are inevitable (and even desirable) in many manufacturing settings (see [24]; [28]; [22]), and typically uses simple analogies and appeals to common sense to support its analyses. While provocative, this qualitative approach is difficult to evaluate objectively. Further, much of the original OPT technology is cloaked in commercial secrecy, further confounding a balanced appraisal of its efficacy.

Another body of literature significant to this paper concerns queueing networks since our model utilizes a Jackson network [17]. (The background literature on the use of queueing networks to model manufacturing systems is vast, so we refer the reader to [6] for a summary.) [19] developed fluid approximations to discrete flow networks and used

their results to define and characterize network bottlenecks. Other work related to ours is that of [20], who chooses capacities for an open Jackson network to minimize expected time in the system subject to a linear budget constraint. While Kleinrock deals with a problem similar to our facilities design problem of section 4, his sojourn-time criterion is different and his use of a budget constraint produces a different solution. [35] subsequently generalizes Kleinrock's model to queueing networks with arbitrary service times, utilizing a heavy traffic approximation of [15]. Wein's simulation results show that the approximation gives good results. Neither of these works directly address the issue of production bottlenecks.

Some work in management of machine capacity has used somewhat more sophisticated network models in their analysis, such as Bitran and Tirupati [3] and Boxma, Kan, and van Vliet [4]. This research differs from these two papers in several respects. First, in these papers there is either a target level of WIP given exogenously, with the cost of the machines being minimized, or WIP is minimized for a given budget of either capital or capacity. In this paper, there are no exogenously given constraint; rather, our results are more of a "natural" economic equilibrium, that is, what the rational levels of both capital investment in capacity and in WIP would be in the absence of these exogenously given targets. A second difference is that, with the simpler Jackson network model, we are able to give explicit forms of the solutions. This is important to our study, not because of the fact of obtaining closed-form results, but because of the insights we can obtain, particularly related to the important concept of the economic bottleneck. The above mentioned papers provide algorithms for the solutions, rather than analytic expressions. Finally, the prior research along these lines has not focused on the *bottleneck* issue, that is, providing means of identifying the workcenters that are constraining flows. In this paper, the economic bottleneck concept allows the framing of capacity and demand decisions in terms of identifying which workcenters constrain cash flows, rather than the more traditional product-flow bottlenecks.

Bottlenecks are also an important topic in the control of production systems, as evidenced by the increasing research interest in bottleneck management. Adams, Balas, and Zawack [1] developed a shifting bottleneck algorithm which iteratively scheduled the bottleneck resource in a classic jobshop scheduling problem to minimize makespan. Chang, Matsuo, and Sullivan [7] examined the makespan scheduling of a flexible manufacturing system, and used the alternative routings available with an FMS to develop a beam-search solution procedure which scheduled around bottleneck machines. Pence, Meegeath, and Morrell [27] describe a cycle-scheduling algorithm for scheduling a single bottleneck when set-up times are sequence dependent.

While this summary illustrates the growing research interest in the management of production bottlenecks, there has been relatively little investigation of the economic consequences of production bottlenecks. Banker, Datar, and Kekre [2] demonstrate how production bottlenecks can affect cost accounting and pricing decisions in stochastic production systems. Morton and Singh [25] generalize these results to deterministic systems and develop a resource pricing methodology based on the busy-period of the resource. Morton, Lawrence, Rajagopolan, and Kekre [26] used a related busy-period resource pricing methodology in the development of a production scheduling algorithm. However, none of these papers has explicitly examined the economic consequences of production bottlenecks, a contribution of this paper.

3. GENERAL MODEL

The general problem we address is that of determining capacities for workcenters and determining the demand rates for each product manufactured in a facility such that long-run average profits are maximized. We subsequently show that production bottlenecks naturally arise from this optimization.

Consider a production facility which manufactures n distinct products (or product classes) with product j having demand rate λ_j items per unit time. These n products are processed in one or more of m workcenters, each with production capacity of μ_k units per unit time. Workcenter capacities μ_k and demand rates λ_j are thus the decision variables of this problem; let $\mu \triangleq (\mu_1, \ldots, \mu_m)$ denote the vector of processing rates and $\lambda \triangleq (\lambda_1, \ldots, \lambda_n)$ denote the vector of demand rates. The sequence of processing in the facility can be random (a job shop), sequential (a flow shop), or mixed.

Relevant costs include the period costs $K(\mu)$ of providing capacity at level μ_k for workcenter k and the period flow costs $F(\mu,\lambda)$ arising from congestion. Note that all costs are the per-period (e.g., annualized) expenses incurred using a common time unit. Capacity costs include amortized equipment expense, plant costs, labor, and maintenance expense. Congestion costs include warehousing and storage costs, materials handling expense, insurance costs, inventory tracking and expediting charges, capital opportunity costs, lost customer goodwill, quality expenses arising from deteriorating in-process inventory, and other similar inventory holding expenses. Marginal period profits $M(\lambda)$ consist of product contribution (gross revenue less variable production costs) net marketing expenses associated with producing demand at rate λ. The general form of the profit function $\Pi(\mu, \lambda)$ can thus be written as

$$\Pi(\mu, \lambda) = M(\lambda) - F(\mu, \lambda) - K(\mu). \tag{1}$$

The firm's objective is to maximize the long-run or steady-state average profit per unit time; that is, to maximize $\Pi(\mu, \lambda)$. Implicit in this formulation is the assumption that the facility can, and does, reach steady state.

The focus of this paper is not the solution of (1) *per se*; rather, it is to investigate the consequences of capacity and demand decisions on the congestion in the facility as exhibited by the bottlenecks. Therefore, we concentrate on the special cases obtained by finding the optimal capacities for given demand, and then on the optimal demand for given capacities. The former problem we call the *Facilities Design* problem, and the latter the *Demand Planning* problem. The joint problem of simultaneously determining optimal capacity and demands is an important and difficult one, but is beyond the scope of the present paper. The reader is referred to [5] which addresses this joint problem in a somewhat different context.

3.1. Jackson Network Model

To investigate the problem outlined above, it is necessary to specify more precisely the characteristics of the production facility. We do this by modeling the shop as a Jackson network consisting of m single-machine workcenters [17]. The Jackson network is very flexible and can model a wide range of situations from a pure job shop to a flow line.

We adopt the standard Jackson network assumptions (see, for example, [18]). Each work center k has a service time distribution that is exponential with rate μ_k. We thus assume that the processing rates of all products j have the same mean processing time $1/\mu_k$ on machine k. Jobs arrive to the workcenters as independent Poisson processes, with λ_k the mean arrival rate to workcenter k, and are distinguished solely by the workcenter at which they first arrive. Our model can thus accommodate as many products or product classes as there are workcenters—additional products can be included by the addition of dummy workcenters.

Upon completion of processing at workcenter i, a job moves to workcenter j with (known) probability P_{ij} and leaves the shop with probability $P_{i,m+1}$. The routing matrix $P = (P_{ij})$ is thus an $(m + 1) \times (m + 1)$ stochastic matrix. The matrix P is assumed to be irreducible with $P_{i,m+1} > 0$ for at least one i. Denote the $m \times m$ submatrix of in-shop routing probabilities by P_0. The aggregate arrival rate vector γ is given by $\gamma = \lambda (I - P_0)^{-1}$ where $\lambda = (\lambda_1, \ldots, \lambda_m)$ is the arrival rate vector (we let $\lambda_j \equiv 0$ if there is no job whose initial workcenter is j). The condition for steady state is $\gamma_k < \mu_k$ for all workcenters k. In this case, the probability of n jobs at the workcenters is identical to that for independent *M/M/*1 queues with arrival rates $\gamma_1, \ldots, \gamma_m$ and service rates $\mu_1, \ldots, \mu_m$. This is a result for the *marginal* probabilities of the number of customers in queue only; the workcenters do not behave dynamically as independent *M/M/*1 queues (see [11]).

To specify the congestion cost function $F(\mu, \lambda)$, we assume that it can vary by workcenter, but is proportional to sojourn times at each workcenter. Denote the unit flow cost at workcenter k by F_k, where F_k is the holding or flow cost per unit time for a customer at station k. The long-run average flow cost per unit time is given by (see Appendix):

$$F(\mu, \lambda) = \sum_{k=1}^{m} \frac{F_k \gamma_k}{\mu_k - \gamma_k}. \tag{2}$$

The profit function in (1) is thus

$$\Pi(\mu, \lambda) = M(\lambda) - \sum_{k=1}^{m} \frac{F_k \gamma_k}{\mu_k - \gamma_k} - K(\mu). \tag{3}$$

A critical assumption in our Jackson network model is that changes in arrival rates λ_j do not affect routing matrix P. This is not a problem under the assumptions of an ideal Jackson network, but may not strictly hold when an actual production facility is approximated using a Jackson network. In the latter case, changes in λ may necessitate changes in P which in turn greatly complicates marginal analysis of $\Pi(\mu, \lambda)$ with respect to λ. Simulation experiments on shops having jobs with fixed routings have demonstrated the robustness of our assumption, but we acknowledge this is a potential limitation of our model.

A second assumption of the model is that different products can be distinguished by their initial workcenter. Our modeling objective is to capture the impact of different products on flow costs, which will be related to the congestion induced by the arrival of

jobs to the facility. For a given arrival rate vector λ, the expected flowcosts for a job arriving at workcenter k will be $\Sigma_k F_k \gamma_k/(\mu_k - \gamma_k)$. Since we are concerned with the capturing impact of flowcosts, we may sometimes consider products to be "identical" if they have comparable value, have similar routings, and require approximately the same processing times. Note that we can introduce dummy workcenters at the beginning of processing to further distinguish the different products.

4. FACILITIES-DESIGN PROBLEM

The *Facilities Design problem* is that of determining capacities μ such that the profit in Expression (3) is maximized for given (fixed) demand rate vector λ. We use the term "design" because the capacity decisions for a shop or facility are typically made during the design stage. We model the cost of providing capacity μ_k at workcenter k as a continuous linear function of capacity, so that the aggregate cost per unit time $K(\mu)$ is given by $K(\mu) = \Sigma_k K_k\mu_k$, where K_k is the cost (per unit time) of providing one unit of capacity at workcenter k. Since demand is fixed, maximizing the profit in (3) is equivalent to minimizing costs; that is, the facilities design problem amounts to minimizing

$$C(\mu) = \sum_k K_k\mu_k + \sum_k \frac{F_k\gamma_k}{\mu_k - \gamma_k}. \tag{4}$$

Our use of a linear capacity cost function is consistent with previous literature (e.g., [16]; [20]; [3]; [35]) and is justified as follows. While it is unlikely that an actual capacity cost function will be linear over its entire range, a linear model can closely approximate a "true" cost function in the neighborhood of any chosen capacity μ and so will be a good representation of capacity costs in the relevant range of near-optimal capacities. It is also unlikely that actual capacity cost functions will be continuous since capacity is typically acquired in discrete units. A linear function is therefore an approximation, but a useful one, since it allows marginal analysis to be conducted resulting in important managerial insight, as we subsequently demonstrate.

4.1. Optimal Workcenter Capacities

Since $C(\mu)$ is convex in μ (see Appendix), there exists a unique optimal vector of capacities which minimizes $C(\mu)$. Optimal workcenter capacities can therefore be found by solving the first-order conditions:

$$\frac{\partial C}{\partial \mu k} = -F_k \frac{\gamma_k}{(\mu_k - \gamma_k)^2} + K_k, \equiv 0 \tag{5}$$

and are given by

$$\mu_k^* = \gamma_k + \sqrt{\frac{F_k\gamma_k}{K_k}}. \tag{6}$$

Expression (6) is a network generalization of Hillier's result for a single server *M/M/*1 queue [16]. If $\lambda_j > 0$ for one j, then $\gamma_k > 0$ for all k, by the irreducibility of the Jackson network. Thus, the optimal capacities μ_k^* are strictly positive and are equal to aggregate arrival rate γ_k plus a capacity cushion $\sqrt{F_k\gamma_k/K_k}$. The capacity cushion at one workcenter is therefore independent of cost parameters K_k and F_k of other workcenters—each workcenter can be considered in isolation from other centers. This property is a consequence of the product-form solution to the Jackson network—in networks without a product-form solution, we would anticipate a greater interaction between the various workcenters.

The optimal capacity utilization for workcenter k, $\rho_k^* \triangleq \gamma_k/\mu_k^*$, is obtained from (6):

$$\rho_k^* = \frac{\sqrt{K_k\gamma_k}}{\sqrt{K_k\gamma_k}+\sqrt{F_k}} = \left(1 + \sqrt{\frac{F_k}{K_k\gamma_k}}\right)^{-1} \tag{7}$$

Flow shops and pure job shops represent two special cases of this general result.

Flow Shop. For a flow shop, the routing is deterministic and each job follows the same sequence of workcenters. All jobs begin at workcenter 1 and proceed in ascending order through workcenter m, after which they leave the system. The routing matrix P thus consists of ones immediately above the diagonal and zeros elsewhere. Since all jobs start at workcenter 1, there is only one job type which arrives at rate λ_1, so we may write the optimal work center capacities and utilizations as

$$\mu_k^* = \lambda_1 + \sqrt{\frac{F_k\lambda_1}{K_k}} \qquad \rho^* = \left(1 + \sqrt{\frac{F_k}{K_k\lambda_1}}\right)^{-1} \tag{8}$$

which is identical to the single-server *M/M/*1 result of Hillier [16].

Pure Job Shop. A Jackson network model of a pure job shop assumes that from any workcenter a job may go to any workcenter (including the one it just left) or out of the shop with equal probability and that the exogenous arrival rate of jobs to each workcenter is identical. Thus, the routing sub-matrix P_0 consists of identical entries, each of which is $1/(m+1)$. The matrix $(I - P_0)^{-1}$ consists of 2's on the diagonal and 1's off the diagonal. The aggregate arrival rate γ_k at each workcenter k is therefore $\gamma_k = (m+1)\lambda_0$, where λ_0 is the common exogenous arrival rate to each workcenter. The optimal capacity and utilization of workcenter k are

$$\mu_k^* = \lambda_0(m+1) + \sqrt{\frac{\lambda_0(m+1)F_k}{K_k}} \qquad \rho^* = \left(1 + \sqrt{\frac{F_k}{\lambda_0(m+1)K_k}}\right)^{-1}, \tag{9}$$

a close analog of the [16] single-server result.

4.2. Designing Optimal Bottlenecks

From expression (7) for ρ_k^*, it is clear that optimal workcenter utilizations will seldom be equal—this occurs only in the unlikely event that two or more ratios $F_k/K_k\gamma_k$ are equal. By our earlier definition of a production bottleneck as the workcenter with the largest

utilization ρ_k, this result demonstrates that production bottlenecks are the usually inevitable outcome of cost minimization. Bottlenecks therefore do not represent a failure in facilities design, but rather arise naturally from the optimizing efforts of design and production personnel.

Expression (7) offers additional insights. First, all else equal, as capacity costs K_k increase, utilization levels ρ_k^* increase—the more expensive is workcenter capacity, the more highly utilized it will be and the more likely it will be a bottleneck. Conversely, a workcenter with large flowcosts F_k will have a reduced capacity utilization and will less likely be a bottleneck. Finally, as job arrival rate γ_k increases (all else equal), the greater will be capacity utilization—the busier the workcenter, the greater is the likelihood that it is a bottleneck. These outcomes are intuitively satisfying. For *both* pure flow shops and pure job shops the ratio K_k/F_k is sufficient to determine the production bottleneck, which is the workcenter with the smallest ratio. For pure job shops and pure flow shops with equal unit costs of congestion and capacity, the optimal design is completely balanced production.

Bottleneck Sensitivity. We now examine the effect of changes in the parameters on the optimal utilizations ρ_k^*. Differentiating (7) with respect to λ_j, F_k, and K_k gives

$$\frac{\partial \rho_k^*}{\partial \lambda_j} = \frac{R_{jk}}{2(\rho_k^*)^2} \sqrt{\frac{F_k}{K_k \gamma_k^3}} > 0 \tag{10}$$

$$\frac{\partial \rho_k^*}{\partial F_k} = -\frac{(\rho_k^*)^2}{2\sqrt{F_k K_k \gamma_k}} < 0 \tag{11}$$

$$\frac{\partial \rho_k^*}{\partial K_k} = \frac{(\rho_k^*)^2}{2} \sqrt{\frac{F_k}{K_k^3 \gamma_k}} > 0 \tag{12}$$

where $R_{jk} \triangleq (I - P_0)^{-1}{}_{jk}$ is the mean number of times workcenter k is visited by a job initiating service at workcenter j [19].

Thus, ρ_k^* is increasing in λ_j and K_k, but decreasing in F_k. From (10), increasing *any* λ_j will result in an increase in *all* ρ_k^*'s for which $R_{jk} > 0$—this change will be proportional to the expected number of times the job visits workcenter k. Changing λ_j will have the greatest impact on non-bottleneck workcenters (those with relatively small utilizations) which are visited relatively often—increases in λ_j will disproportionately increase the utilization of these centers. From (11) and (12), changing F_k or K_k affect the optimal utilization of workcenter k only. The impact is greatest for bottleneck workcenters (those with large utilizations) and for those with relatively low aggregate arrival rates γ_k. Increasing flow costs at a workcenter has the effect of *decreasing* capacity utilization, while increasing the costs of capacity *increase* the utilization at that workcenter alone.

Sensitivity of Costs. We now examine the sensitivity of total costs $C(\mu^*)$ to changes in workcenter parameters. Such information is managerially valuable since it focuses attention on workcenters with the largest potential return on improvement. Substituting the optimal capacities μ^* into (3), we have the minimum cost associated with the optimal design:

$$C(\mu^*)=2\sum_{k=1}^{m}\sqrt{F_k K_k \gamma_k}+\sum_{k=1}^{m} K_k \gamma_k \,. \tag{13}$$

Observe that the cost at the optimum grows linearly in K_k, but as the square root of F_k. Examining the impact of changing the various parameters on the optimal cost, from (4) we immediately have

$$\frac{\partial C(\mu^*)}{\partial K_k} = \mu_k^* = \frac{\gamma_k}{\rho_k^*} \tag{14}$$

This result indicates that for fixed demand, cost-reduction efforts (perhaps through improvements in technology or redesign) should focus on those workcenters with the greatest assigned capacities μ_k^*. Note that these workcenters may well not be those with largest capacity costs K_k, or those with the greatest utilization ρ_k, as intuition might first suggest. Indeed, all else equal, reducing the capacity costs of *non-bottleneck* workstations (those with lower utilizations) will provide faster paybacks than will reducing costs at bottleneck stations.

Considering changes in flow costs F_k at workcenter k, we have

$$\frac{\partial C(\mu^*)}{\partial F_k} = \frac{\rho_k^*}{1-\rho_k^*} = \sqrt{\frac{K_k \gamma_k}{F_k}} \,. \tag{15}$$

Here, the largest rate of cost reduction is obtained by reducing F_k at the workstation with the highest capacity utilization, as intuition would suggest. More surprising is that, all else equal, the best total cost improvement rate occurs at that workcenter with the *lowest* value of F_k, since this corresponds to workcenters with low utilizations in the optimal solution.

Finally, consider changes in total cost due to increases in λ_j (the arrival rate of job type j), perhaps due to changes in marketing effort. From (6) and (13) we have:

$$\frac{\partial C(\mu^*)}{\partial \lambda_j} = \sum_{k=1}^{m} R_{jk}\left(K_k + \sqrt{\frac{F_k K_k}{\gamma_k}}\right) = \sum_{k=1}^{m} R_{jk} K_k / \rho_k^* \tag{16}$$

Recalling that R_{jk} is the mean number of visits of product j to workstation k, we see that, other things equal, products which visit more costly workstations more frequently produce larger cost increases when demand increases. Similarly, products which visit workstations more frequently with higher flowcosts also produce larger cost increases in response to increased demand for those products. Although these two results are quite intuitive, note that (16) provides a means of explicitly evaluating the impact of increasing demand for each product on costs.

4.3. Cost of Balancing Capacity

A frequently cited goal of production managers is to balance production, meaning that all machines and workcenters are equally utilized, thus "eliminating" bottlenecks. The intent of such a policy is to ensure that capital equipment and fixed labor is not underutilized, thereby incurring supposed opportunity costs from lost production.

A policy of balanced production is equivalent to insisting that traffic intensities $\rho_k = \gamma_k/\mu_k$ be identical for all machines k. Let ρ be this common traffic intensity and C_B the cost associated with the balanced utilization policy. Cost function (4) becomes:

$$C_B(\rho) = \sum_k F_k \frac{\rho}{1-\rho} + \sum_k \frac{K_k \gamma_k}{\rho} \tag{17}$$

The optimal (common) utilization level can be found by solving the first-order condition

$$C_B'(\rho) = \frac{1}{(1-\rho)^2} \sum_k F_k - \frac{1}{\rho^2} \sum_k K_k \gamma_k \equiv 0, \tag{18}$$

yielding

$$\rho^* = 1 - \frac{\sqrt{\Sigma_k F_k}}{\sqrt{\Sigma_k F_k} + \sqrt{\Sigma_k K_k \gamma_k}}. \tag{19}$$

The cost associated with this balanced policy is

$$C_B(\rho^*) = 2\sqrt{\left(\sum_k K_k \gamma_k\right)\left(\sum_k F_k\right)} + \sum_k K_k \gamma_k. \tag{20}$$

As with $C(\mu^*)$, $C_B(\rho^*)$ grows linearly in K_k and as the square root of F_k. The difference between $C(\rho^*)$ and $C_B(\mu^*)$ is

$$C_B(\rho^*) - C(\mu^*) = 2\left(\sqrt{\left(\sum_k K_k \gamma_k\right)\left(\sum_k F_k\right)} - \sum_k \sqrt{F_k K_k \gamma_k}\right). \tag{21}$$

This non-negative (and typically positive) quantity represents the additional costs that the firm must pay for adhering to a balanced production policy.

In summary, the results of this section show that cost-minimizing behavior leads to the deliberate creation of production bottlenecks. Bottlenecks are thus an inevitable result of the optimization of resource capacities. Rather than attempting to balance production or eliminate bottlenecks, production managers should in fact *design* bottlenecks into the production facility, with the location of these bottlenecks determined by appropriate cost (or profit) considerations.

5. DEMAND-PLANNING PROBLEM

Our analysis in the previous section addressed the design of productive capacity with demand rates assumed to be fixed. In this section we consider the analogous *Demand*

Planning problem: determining optimal demand rates for given *fixed* capacity. This applies to situations in which management desires to maximize profits at an existing production facility by adjusting demand levels to match the capabilities of an existing plant. We assume that demand levels can be modified by the firm through adjustments to advertising expenditures, price incentives, sales effort and focus, and similar marketing initiatives, but note that the marketing initiatives themselves are not decision variables in our model. We further assume that the plant is currently operating in the region of optimal demand levels, so that required demand changes are relatively small.

Capacity costs $K(\mu)$ are now sunk, and so are irrelevant, whereas marginal revenues $M(\lambda)$ are now relevant. Since production capacities μ_k are fixed, the relevant range of feasible production volumes λ_k is relatively narrow. Consequently, we model marginal revenues as a linear function of production volumes: $M(\lambda) \triangleq \sum_{j=1}^{n} M_j \lambda_j$. As with the design problem of section 3, we distinguish products by the workcenter at which they initiate service. If there are no products with workcenter j as the initial workcenter, we set $M_j = 0$, effectively creating a "dummy" product. Clearly these dummy products will have $\lambda_j^* = 0$, since they only add flow costs and do not contribute to the profitability of the operation.

Taken together with congestion costs, the objective of the demand planning problem is to maximize contribution $\Pi(\lambda)$ defined as marginal revenues less congestion costs:

$$\Pi(\lambda) \triangleq \sum_{j=1}^{m} M_j \lambda_j - \sum_{k=1}^{m} \frac{F_k \lambda_k}{\mu_k - \lambda_k}. \tag{22}$$

The demand-planning problem is complicated by the fact that γ is a function of λ and we must therefore explicitly consider both non-negativity of λ as well as feasibility of γ. This latter condition arises because the traffic intensity at each workcenter must be less than unity (i.e., $\rho_k \triangleq \mu_k/\gamma_k < 1$). Thus, the model for *demand planning* becomes

$$\begin{aligned} &\max \Pi(\lambda) \\ &\text{subject to} \\ &\lambda \geq 0 \\ &\lambda(I - P_0)^{-1} < \mu \end{aligned} \tag{23}$$

The first constraint insures non-negative arrivals, while the second ensures the feasibility of γ.

5.1. Properties of the Solution

Program (23) can be solved using standard nonlinear programming techniques, but it will be useful to establish several of its properties, including the existence of a solution. Since the set $S_2 = \{\lambda \in R^+: \lambda(I - P_0)^{-1} < \mu\}$ is convex and Π is convex on S_2, there is a Kuhn-Tucker point (λ^*, θ^*), where θ is the vector of dual variables associated with the non-negativity constraints (see Appendix). Elements θ_j of the dual vector θ represent the

reduced margins of the various products available for production (analogous to reduced costs in linear programming). For products being produced ($\lambda_j > 0$), the reduced margin is zero ($\theta_j = 0$). For products not being produced ($\lambda_j = 0$) the reduced margins are non-negative ($\theta_j \geq 0$), representing the amount by which margin M_j must be increased in order to make product j attractive enough for production.

If revenues M_j are not sufficient to cover costs for product j then it should not be produced, and if this is true for all products then it is optimal not to produce at all. The following conditions indicate when it is not profitable for the firm to produce a given product: (1) If $M_j - \sum_{k=1}^{m} F_k R_{jk}/\mu_k \leq 0$ for all j, then the product initiating service on machine j should not be produced (i.e., $\lambda^* = 0$); (2) Conversely, if $M_j - \sum_{k=1}^{m} F_k R_{jk}/\mu_k > 0$ for some j then $\lambda_i^* > 0$ for some i (see Appendix).

Henceforth, to avoid the no-production situation ($\lambda_j^* = 0$ for all j), we will assume that $M_j - \sum_{k=1}^{m} F_k R_{jk}/\mu_k > 0$ for at least one j. By Corollary 28.3.1 of [29], λ^* maximizes $\Pi(\lambda)$, and (λ^*, θ^*) satisfy the first-order conditions

$$M_j + \theta_j - \sum_{k=1}^{m} \frac{F_k \mu_k R_{jk}}{(\mu_k - \gamma_k)^2} \equiv 0, \qquad j = 1, \ldots, m \tag{24}$$

Thus, γ^* may be expressed in terms of θ^* by solving (24):

$$\gamma_k^* = \mu_k - \sqrt{\frac{F_k \mu_k}{\sum_{i=1}^{m} Q_{ki}(M_i + \theta_i^*)}} = \mu_k - \sqrt{\frac{F_k \mu_k}{\hat{M}_k + \hat{\theta}_k^*}}. \tag{25}$$

in which $\hat{M} = QM$, $\theta = Q\theta$, and $Q = I - P_0$.

Expanding their definitions, we have $\hat{M}_k = (1 - P_{kk})\, M_k - \sum_{i \neq k} P_{ki} M_i$ and $\hat{\theta}_k^* = (1 - P_{kk})\theta_k^* - \sum_{i \neq k} P_{ki}\, \theta_i^*$ respectively. Thus, the values $\hat{M}_k$ and $\hat{\theta}_k^*$ have the following economic interpretations. $\hat{M}_k$ is the expected margin obtained by incrementing the arrivals of product k by one additional unit, less the weighted opportunity costs of margins foregone at downstream workcenters—$\hat{M}_k$ can be positive, zero, or negative. Similarly, $\hat{\theta}_k^*$ is the reduced margin at center k, less the weighted reduced margins at downstream centers, and can be positive, negative, or zero. Since θ_j is positive only for a workcenter j with $\lambda^* = 0$, it represents the additional opportunity cost of depriving *other* jobs from using the capacity of workcenter j. Thus $(\hat{M}_k + \hat{\theta}_k^*)$ represents the aggregate benefits of increasing arrivals of product k by one additional unit.

From (25) we have

$$\lambda_j^* = \sum_{k=1}^{m} Q_{kj} \left(\mu_k - \sqrt{\frac{F_k \mu_k}{(\hat{M}_k + \theta_k^*)}} \right) \tag{26}$$

so the optimal utilization ρ_k^* at workcenter k is

$$\rho_k^* = 1 - \sqrt{\frac{F_k}{\mu_k(\hat{M}_k + \hat{\theta}_k^*)}}. \tag{27}$$

The bottleneck workcenter will be the one for which $\sqrt{F_k/\mu_k(\hat{M}_k + \hat{\theta}_k^*)}$ is smallest. Workcenters with large flow costs or small capacities (i.e., small μ_k) will tend to have small capacity cushions and therefore be production bottlenecks. One again, we see that production bottlenecks are the inevitable outcome of optimizing behavior.

Although θ^* cannot be expressed in closed form, a simple condition indicates whether some will be positive: if $\hat{M}_k \leq F_k/\mu_k$, then $\theta_k^* > 0$, and hence $\lambda_k^* = 0$ (see Appendix). To interpret this condition, observe that F_k/μ_k is the average flow cost accumulated during the processing of job k on its first work center, k. The value of $\hat{M}_k = M_k - \sum_j P_{kj} M_j$ can be interpreted as follows. After processing on workcenter k, the job goes to workcenter j with probability P_{ij}. That workcenter thus has an opportunity cost of M_j for processing a job of type k, since no revenue is received. Thus, $\hat{M}_k$ can be interpreted as the marginal revenue net the expected opportunity cost of processing on the second workcenter. If this is not sufficient to meet the expected flow costs during processing (which is a lower bound on flow costs), then it will clearly not be profitable to take type k jobs.

The planning problem is complicated by the possibility of the optimal solution not being an interior point, that is, having $\lambda_j^* = 0$ for some j's. However, if we *do* have an optimal point in the interior, it is of the form:

$$\gamma_k^* = \mu_k - \sqrt{\frac{F_k \mu_k}{\hat{M}_k}} \tag{28}$$

and

$$\lambda_j^* = \sum_{k=1}^{m} \left(\mu_k - \sqrt{\frac{F_k \mu_k}{\hat{M}_k}} \right) Q_{kj}, \tag{29}$$

with optimal utilization

$$\rho_k^* = 1 - \sqrt{\frac{F_k}{\hat{M}_k \mu_k}} \tag{30}$$

Equation (29) is, again, the network analog of Hillier's [16] *M/M*/1 single-server results.

5.2. Special Cases

Flow Shop. The profit function for planning problem in a flow shop is:

$$\Pi(\lambda_1) = M\lambda_1 - \sum_{k=1}^{m} \frac{F_k \lambda_1}{\mu_k - \lambda_1} \tag{31}$$

with corresponding first-order condition

$$R'(\lambda_1) = M - \sum_{k=1}^{m} \frac{F_k \mu_k}{(\mu_k - \lambda_1)^2} \equiv 0. \tag{32}$$

There will be a positive optimal λ_1 iff $M > \sum_{k=1}^{m} F_k/\mu_k$, and the production bottleneck will be the workcenter with the smallest capacity μ_k.

Uniform Utilization. Unlike the capacity design problem, it may not be possible to have the same utilization at all workcenters for the planning problem. However, the following two conditions are necessary and sufficient for uniform utilization to be possible (see Appendix): (1) $\mu_j > \sum_k \mu_k P_{kj}$ for all workcenters j with an exogenous arrival stream, and (2) $\mu_j = \sum_k \mu_k P_{kj}$ for all workcenters j without an exogenous arrival stream. For those cases in which equal utilization *is* possible, we can determine the optimal utilization ρ^*. We obtain the profit associated with balanced utilization, Π_B in a manner similar to section 3:

$$\Pi_B(\rho) = \rho \sum_{k=1}^{m} \hat{M}_k \mu_k - \frac{\rho}{1-\rho} \sum_{k=1}^{m} F_k . \tag{33}$$

The optimal utilization for balanced production is therefore

$$\rho^* = 1 - \sqrt{\frac{\Sigma_k F_k}{\Sigma_k \mu_k \hat{M}_k}} \tag{34}$$

We may then obtain similar results as for the facilities design problem with equal utilization.

6. ECONOMIC BOTTLENECK

Our analysis of the facilities design and demand planning problems has shown that bottlenecks naturally arise when firms organize capacity design and demand volumes to minimize costs and maximize profits. Implicit in this analysis is the notion of an *economic bottleneck*, defined to be that workstation which most severely increases costs or limits profits. Below we offer a formal definition of an economic bottleneck, distinguish it from a production bottleneck, and demonstrate that the two do not necessarily coincide.

Consider a production facility with existing capacity levels μ and fixed production demands λ. Management of the firm wishes to reduce costs and improve throughput rates by increasing capacity at the appropriate workstation(s). Capacity might be increased by incremental technological improvements, better management of secondary resources such as labor and tooling, reduced setup times, and so forth. We distinguish these types of continuous improvement activities from the initial design of the facility. Given limited resources, at which workstation(s) should management focus its attention, and what are the potential benefits?

Under these assumptions, initial capacity costs $K(\mu)$ are sunk (since the plant is in operation), and contribution margins $M(\lambda)$ are constant (since production demands λ are fixed). Consequently, the objective function facing the firm is simply to minimize congestion costs

$$F(\mu) = \sum_{k=1}^{m} \frac{F_j \gamma_j}{\mu_j - \gamma_j} \tag{35}$$

with respect to μ. We define the *economic bottleneck* to be that station for which marginal increases in capacity provides the largest decrease in congestion costs; that is, the workcenter k for which

$$\frac{\partial F(\mu)}{\partial \mu_k} = -\frac{F_k \gamma_k}{(\mu_k - \gamma_k)^2} \tag{36}$$

is a minimum. This definition of an economic bottleneck provides the marginal benefit or shadow price of marginal increases in capacity at a particular workcenter. This result applies to any production facility, whether or not it is optimally configured with regard to λ and μ. Note also that, as defined in (36), the costs of additional capacity are not explicitly considered.

If the facility is optimally planned, that is $\lambda = \lambda^*$, then further results can be obtained. By the Envelope Theorem (see, for instance, [34]), $\partial F(\lambda_k^*)/\partial \mu_k = \partial F(\lambda_k)/\partial \mu_k|_{\lambda_k^*}$, so that

$$\frac{\partial F}{\partial \mu_k} = -\frac{F_k \gamma_k^*}{(\mu_k - \gamma_k^*)^2} = -(\hat{M}_k + \hat{\theta}_k^*)\, \rho_k^* \tag{37}$$

If $\lambda_j^* > 0$ for all j, then $\hat{\theta}^* = 0$ and $\partial F/\partial \mu_j = -\hat{M}_k \rho_k^*$.

Expression (37) connects the concept of an economic bottleneck to that of a production bottleneck. An economic bottleneck combines the definition of a production bottleneck (via utilization ρ_k^*), with the economic benefit $(\hat{M}_k + \hat{\theta}_k^*)$ of increasing the capacity of workstation k. Increasing capacity has the effect of allowing additional units of product k to be accepted for processing with net benefit $\hat{M}_k$, and simultaneously allows additional units of other products $i \neq k$ to be processed with net benefit $\hat{\theta}_k^*$.

The economic bottleneck concept has important consequences for the improvement of facilities. While the *production* bottleneck is the workstation which constrains capacity, the *economic* bottleneck is the workcenter which constrains profitability. Clearly the two need not be identical and, indeed, will typically be different. Indeed, Expression (37) shows that the concepts coincide only when $\hat{M}_k + \hat{\theta}_k^*$ are identical for all workcenters k, an unlikely occurrence. Capacity decisions based on production bottleneck considerations may therefore not be the most profitable ones, as shown in the following example.

7. NUMERICAL EXAMPLE

A simple example will serve to illustrate several of our results. Consider a production facility with 5 single-server workstations having routing matrix given in Table I, and with other parameters given in Table II (the base case).

The facility design problem is straightforward, as discussed previously, since the optimal capacities will always be positive and given in closed form, for linear cost of capacity functions. Using the base case values for F, K, and demand arrivals λ as shown in Table III, equation (6) gives the optimal capacities μ^* in Table III.

For the demand planning problem, using base case values for F, M, and μ as given in Table III, equation (26) gives optimal values of λ^* as in Table III. Since all λ_k^* are positive,

Table I Five-Workstation Example: Routing Matrix

From	To Workstation					
Workstation	1	2	3	4	5	Out
1		1/3	2/3			
2	1/6		1/3	1/3		1/6
3	1/6	1/6		1/3		1/3
4		1/6	1/6		1/3	1/3
5			1/3	1/3		1/3

Table II Five-Workstation Example: Economic Parameters

	Workstation				
	1	2	3	4	5
F	1.0	5.0	20.0	3.0	10.0
M	23.0	20.0	24.0	13.0	15.0
K	1.0	4.0	7.0	0.5	0.5

Table III Base Case and Optimal Workcenter Parameters

		Workstation				
		1	2	3	4	5
Base	μ	46.0	56.0	111.0	85.0	50.0
Case	λ	17.0	8.0	28.0	15.0	10.0
	ρ	0.90	0.88	0.87	0.88	0.70
Facility	μ^*	47.7	57.1	112.7	96.4	61.5
Design	ρ^*	0.87	0.86	0.85	0.78	0.57
	λ^*	10.1	8.8	36.0	5.0	14.5
Demand	ρ^*	0.74	0.85	0.88	0.77	0.73
Planning	$\hat{M}\,\rho^*$	0.2	3.2	11.0	0.5	1.9
	$\theta^*, \hat{\theta}^*$	0.0	0.0	0.0	0.0	0.0

this initial example is "full rank" for the demand problem in that all machines have jobs arriving at the optimum.

For the base case, Table III shows that station 1 is the *production bottleneck* since it has the highest capacity utilizations among the five stations (although stations 2 and 4 could be considered secondary bottlenecks). Conventional analysis therefore suggests that station 1 should first be considered for relief before the other stations. However, determination of the *economic bottleneck* using equation (37) demonstrates that station 3 is the economic bottleneck, with a dual price that is substantially greater than the second largest (station 2) as shown in Table III. To emphasize this result further, Figure 1 shows the impact of increasing or decreasing capacity on profits at each of the workstations, and clearly shows station 3 to be the economic bottleneck. Costs will decline most rapidly by increasing capacity at the economic bottleneck, not the production bottleneck.

Finally, Figure 2 uses equation (17) to compare optimal costs with actual costs balanced production is enforced. For our example, this figure shows that the negative impact of balancing utilization is small when that utilization is optimal, but that there is a stiff penalty if the optimum is missed. Furthermore, as the common utilization approaches unity, the penalty gets increasingly worse. Similar results are obtained when using equation (33) to compare optimal profits with balanced production profits.

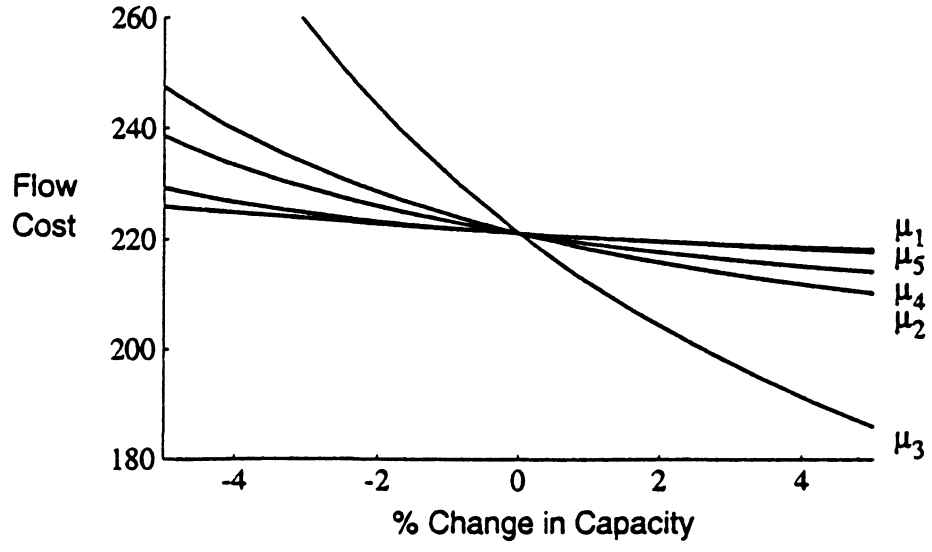

Figure 1 Flow cost versus capacity.

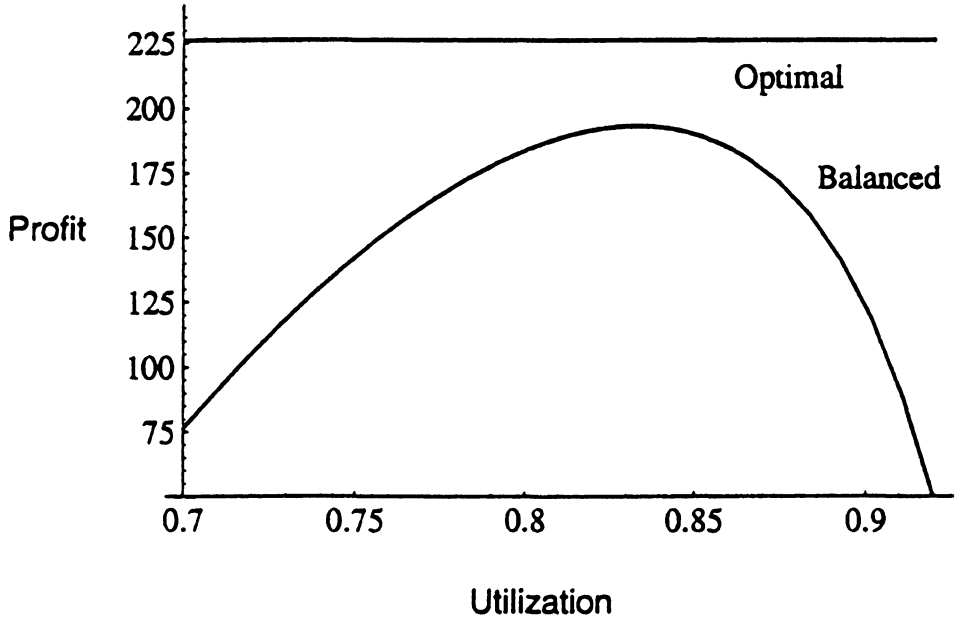

Figure 2 Optimal versus balanced production profits.

8. CONCLUSIONS

In this paper, we have critically analyzed bottlenecks from an economic perspective. Using queueing network theory, we have demonstrated that bottlenecks are inevitable when there are differences in job arrival rates, processing rates, or costs of productive resources. These differences naturally lead to the creation of bottlenecks both when designing production facilities to meet anticipated demand, and when determining demand loads for existing facilities.

Our results have a number of implications for the management of stochastic production facilities:

- Care must be taken when managing a "bottleneck," since several definitions are in current use, each requiring a different management response;
- Production bottlenecks are the natural outcome of minimizing costs when designing production facilities, or of maximizing profits when determining demands loads;
- Balanced production is rarely optimal, from a purely economic perspective;
- The concept of an *economic bottleneck* provides an economic basis for bottleneck management and improvement.

Our treatment of the demand-planning problem demonstrates that consideration of flow costs due to congestion and the demands different products place on facilities can contribute to the unprofitability of certain items. In those cases, our model gives rise to dual variables which measure the degree of unprofitability for those items.

We have found that while the various production bottleneck concepts are related to the economic bottleneck, they do not necessarily coincide. In our opinion, the economic bottleneck is ultimately a more useful concept for managers, since it shows the way to the best financial return on capacity investment. Thus, relieving an economic bottleneck will bring greater returns to the firm than relieving a production bottleneck. The latter will merely increase output by the greatest amount without regard for profitability. Furthermore, we have found that attempting to balance the shop by equalizing utilization is also not economically the best course. However, for the facilities-design problem, our solution can be considered the *economically* balanced shop.

References

1. Adams, J., E. Balas, and D. Zawack, The shifting bottleneck procedure for job shop scheduling, *Management Science* ***34***(1988), 391–401.
2. Banker, R.D., S.M. Datar, and S. Kekre, Relevant costs, congestion and stochasticity in production environments, *Journal of Accounting and Economics* ***10***(1988), 171–197.
3. Bitran, G.R. and D. Tirupati, Tradeoff curves, targeting and balancing in manufacturing queueing networks, *Operations Research* ***37***(1989), 547–564.
4. Boxma, O.J., A.G.H, Rinooy Kan, and M. van Vliet, Machine allocation problems in manufacturing networks, *European Journal of Operational Research* ***45***(1990), 47–54.
5. Buss, A.H., S.R. Lawrence, and D.H. Kropp, volume and capacity interaction in facilities design, *IIE Transactions* ***26***(1994), 36–49.
6. Buzacott J.A., and J.G. Shanthikumar, *Stochastic models of manufacturing systems*, Prentice-Hall, Englewood Cliffs, N.J., 1993.

7. Chang, Y.L., H. Matsuo, and R.S. Sullivan, A bottleneck-based beam search for job scheduling in a flexible manufacturing system, *International Journal of Production Research* **27**(1989), 1949–1961.
8. Chase, R.B. and N.J. Aquilano, *Production and operations management: A life cycle approach*, Fifth Edition, Irwin, Homewood, IL, 1989.
9. Chen, H. and A. Mandelbaum, Discrete flow networks: Bottleneck analysis and fluid approximations, *Mathematics of Operations Research* **16**(1991), 408–446.
10. Chung, K.L. *Markov chains with stationary transition probabilities*, Second Edition, Springer-Verlag, Berlin, 1967.
11. Disney, R.L. and P.C. Kiessler, *Traffic processes in queueing networks: A Markov renewal approach*, Johns Hopkins University Press, Baltimore, 1987.
12. Fogarty, D.W., T.R. Hoffman, and P.W. Stonebraker, *Production and operations management*, South-Western Publishing, Cincinnati, 1989.
13. Goldratt, E.M., and J. Cox, *The goal: Excellence in manufacturing*, North River Press, Croton-on-Hudson, N.Y., 1984.
14. Goldratt, E.M., *Theory of constraints*, North River Press, Croton-on-Hudson, N.Y., 1990.
15. Harrison, J.M., and R. Williams, Brownian models of open queueing networks with homogeneous customer populations, *Stochastics* **22**(1987), 77–115.
16. Hillier, F.S., Economic models for industrial waiting line problems, *Management Science* **10**(1963), 119–130.
17. Jackson, J.R., Networks of waiting lines, *Operations Research* **5**(1957), 518–521.
18. Kelly, F.P. *Reversibility and stochastic networks*, John Wiley, New York, 1979.
19. Kemeney, J.G., J.L. Schnell, and A.W. Knapp, *Denumerable Markov chains*, Springer-Verlag, New York, 1976.
20. Kleinrock, L. *Communication nets: Stochastic message flow and delay*, Dover Publications, New York, 1964.
21. Krajewski, L.J., and L.P. Ritzman, *Operations management: Strategy and analysis*, Second edition, Addison-Wesley, Reading, MA, 1990.
22. Lambrecht, M.R. and L. Decaluwe JIT and constraint theory: The issue of bottleneck management, *Production and Inventory Management Journal*, **3**(1988), 61–66.
23. Lawrence, S.R., and A.H. Buss, Shifting production bottlenecks: Causes, cures, and conundrums *Productions and Operations Management* **3**(1994).
24. Lundrigan, R., What is this thing called OPT? *Production and Inventory Management* **27**(1986), 2–12.
25. Morton, T.E., and M.R. Singh, Implicit costs and prices for resources with busy periods, *Journal of Manufacturing and Operations Management* **1**(1988), 305–322.
26. Morton, T.E., S.R. Lawrence, S. Rajagopolan, and S. Kekre, Sched-Star: A price-based shop scheduling module, *Journal of Manufacturing and Operations Management* **1**(1988), 131–181.
27. Pence, N.A., J.D. Megeath, and J.S. Morrell, Coping with temporary bottlenecks in a several-stage process with multiple products, *Production and Inventory Management* **31**(1990), 5–6.
28. Plenert, G. and T.D. Best, MRP, JIT, and OPT: What's 'best'? *Production and Inventory Management* **27**(1986), 22–29.
29. Rockafeller, T. *Convex analysis*, Princeton University Press, Princeton, N.J., 1971.
30. Sadowski, R.P. and C.M. Harmonosky, Production and project scheduling, Chapter 3.9 in *Production Handbook*, Fourth Edition, John A. White (ed.), John Wiley, New York, 1987, pp. 3·137–3·175.
31. Schmenner, R.W. *Production/operations management: Concepts and situations*, Fourth Edition, MacMillan Publishing New York, 1990.
32. Schonberger, R.J., and E.M. Knod, Jr. *Operations management: Serving the customer*, Third Edition, Business Publications, Plano, TX, 1988.
33. Umble, M.M. and M.L. Srikanth, *Synchronous manufacturing: Principles for world class excellence*, South-Western Publishing, Cincinnati, 1990.
34. Varian, H.R., *Microeconomic analysis*, Second Edition, W.W. Norton, New York, 1984.
35. Wein, L.M. Capacity allocation in generalized Jackson networks, *Operations Research Letters* **8**(1989) 143–146.

APPENDIX

First, we show that the long-run average flow cost per unit time for the Jackson network defined in section 3 is given by

$$\sum_{k=1}^{m} \frac{F_k \gamma_k}{\mu_k - \gamma_k}. \tag{38}$$

Consider a Jackson Network with m single-server nodes (workcenters) with service rate μ_k at node k. The arrivals to node j form independent Poisson processes, with the mean arrival rate to node j given by λ_j. Job routing is via a Markov transition matrix P, which we assume to be open and irreducible. Each job accumulates a flow cost at rate F_k while at node k. The arrival rate to node j is thus γ_j, which solves the flow balance equations $\gamma = \lambda(I - P_0)^{-1}$, and the system is ergodic providing $\gamma_j/\mu_j < 1$ for each j. Let $X(t)$ be the vector process consisting of the number of jobs at each node and $\pi(n)$ the steady-state probability. Let ψ be a real-valued function on $\mathbf{N}^m$, where $\mathbf{N}$ is the non-negative integers, for which $\Sigma_n\, \psi(n)\pi(n) < \infty$.

The from the ergodicity of $\{X(t)\}$ we have (see [10]):

PROPOSITION 1.

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \psi(X(t))\, dt = \sum_n \psi(n)\pi(n) \quad \textit{almost surely.}$$

Now let $\psi(n) = \sum_{j=1}^m n_j F_j$. Then $\psi(n)$ is the rate at which flow costs are accumulating when there are n_j jobs at node $j, j = 1, \ldots, m$, and $\lim_{T\to\infty} 1/T \int_0^T \psi(X(t))\, dt$ is the long-run average flow cost per unit time for the Jackson network.

PROPOSITION 2. *The long-run average flow cost per unit time in the Jackson network described above is given by*

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \psi(X(t))\, dt = \sum_{k=1}^m \frac{F_k\gamma_k}{\mu_k - \gamma_k}. \tag{39}$$

Proof. By Proposition 1 we have

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \psi(X(t))\, dt = \sum_n \psi(n)\pi(n)$$

$$= \sum_n \sum_{k=1}^m F_k \prod_{j=1}^m \rho_j^{n_j}(1 - \rho_j)$$

$$= \sum_{k=1}^m F_k \sum_{n_k=0}^\infty n_k \rho_k^{n_k}(1 - \rho_k)$$

$$= \sum_{k=1}^m \frac{F_k\rho_k}{1 - \rho_k}$$

$$= \sum_{k=1}^m \frac{F_k\gamma_k}{\mu_k - \gamma_k}.$$

PROPOSITION 3. *$C(\mu)$ is convex on the set $S_1 \triangleq \{\mu: \mu > \gamma\}$.*

Proof. The Hessian of $C(\mu)$ is a diagonal matrix with elements $2F_k\gamma_k/(\mu_k - \gamma_k)^3$, and these are all positive under the assumptions.

PROPOSITION 4. *The profit function* Π *is concave on the set* $S_2 = \{\lambda \in \mathbf{R}^+: \lambda(I - P_0)^{-1} < \mu\}$.

Proof. Let $R = (I - P_0)^{-1}$. Then the second partial derivatives of Π are given by

$$\frac{\partial^2 \Pi}{\partial\lambda_i\,\partial\lambda_i} = -\sum_{k=1}^{m} \frac{2F_k\mu_k R_{ik} R_{jk}}{(\mu_k - \gamma_k)^3} \tag{40}$$

For any non-zero vector $x \in \mathbf{R}^m$,

$$\sum_{i,j} x_i \frac{\partial^2\Pi}{\partial\lambda_i\partial\lambda_j} x_j = -\sum_{k=1}^{m} \frac{2F_k\mu_k}{(\mu_k - \gamma_k)^3} \left(\sum_i x_i R_{ik}\right)\left(\sum_j x_k R_{jk}\right)$$

$$< 0, \tag{41}$$

since $\lambda \in S_2 \Rightarrow \mu_k - \gamma_k > 0$ for all k.

PROPOSITION 5. *There is a Kuhn-Tucker point* (λ^*, θ^*) *in program (23), where* θ *is the vector of dual variables associated with* $\lambda \geq 0$.

Proof. Since the set C is convex with a non-empty interior and the constraints are linear with a feasible λ satisfying the constraints with strict inequality, the program (23) satisfies the conditions of Theorem 28.2 of Rockafeller [29] (appropriately modified for a concave objective) asserting the existence of a Kuhn-Tucker point.

PROPOSITION 6. *If* $M_j - \sum_{k=1}^{m} F_k R_{jk}/\mu_k \leq 0$ *for all j, then the product initiating service on machine j should not be produced (i.e.,* $\lambda^* = 0$*); (2) Conversely, if* $M_j - \sum_{k=1}^{m} F_k R_{jk}/\mu_k > 0$ *for some j then* $\lambda_i^* > 0$ *for some i*

Proof. Follows from the gradient at $\lambda = 0$ and the concavity of Π.

PROPOSITION 7. *If* $\hat{M}_k \leq F_k/\mu_k$, *then* $\theta_k^* > 0$, *and hence* $\lambda_k^* = 0$.

Proof. Suppose, to the contrary, that $\hat{M}_k \leq F_k/\mu_k$, but $\theta_k^* = 0$. Then $\hat{\theta}_k^* \times 0$, since Q is positive only on the diagonal. Furthermore, since $Q_{ki} \leq 0$ for $k \neq i$, $\hat{M}_k \leq F_k/\mu_k$ implies $\gamma_k^* \leq 0$. But $\gamma^* > 0$ under our assumption that $M_j > \sum_k F_k R_{jk}/\mu_k$. Therefore, we must have $\theta_k^* > 0$.

PROPOSITION 8. *Balanced capacity utilization (*i.e., $\rho \triangleq \gamma_j/\mu_j$ *is the same for all workcenters j) is* possible *iff (1)* $\mu_j > \sum_k \mu_k P_{kj}$ *for all workcenters j with an exogenous arrival stream,* and (2) $\mu_j = \sum_k \mu_k P_{kj}$ *for all workcenters j without an exogenous arrival stream.*

Proof. Suppose balanced capacity utilization of $\rho \in (0,1)$ is possible for some vector λ, and let $\gamma = \lambda(I - P_0)^{-1}$. Then $\gamma = \rho\mu$, so $\lambda = \rho\mu(I - P_0)$. Thus, $\lambda_j = \rho\,(\mu_j - \sum_k \mu_k P_{kj})$. If workcenter j has an exogenous arrival stream, then $\lambda_j > 0$, so $\mu_j > \sum_k \mu_k P_{kj}$ must hold.

On the other hand, if workcenter j does not have an exogenous arrival stream, then $\lambda_j = 0$, so we must have $\mu_j = \sum_k \mu_k P_{kj}$.

Conversely, if (1) and (2) hold, then for any $\rho \in (0,1)$, set $\lambda = \rho\mu(I - P_0)$. By (1) $\lambda_j > 0$ for all workcenters with an exogenous arrival stream and by (2) $\lambda_j = 0$ for all workcenters without an exogenous arrival stream. Thus, λ is a feasible vector of arrival rates. Since $\gamma = \rho\mu$, this choice of λ gives equal capacity utilization ρ at all workcenters.