MANAGEMENT SCIENCE
Vol. 53, No. 3, March 2007, pp. 408–420
ISSN 0025-1909 | EISSN 1526-5501 | 07 | 5303 | 0408

informs

DOI 10.1287/mnsc.1060.0636
© 2007 INFORMS

Obtaining Fast Service in a Queueing System via Performance-Based Allocation of Demand

Gérard P. Cachon
Operations and Information Management, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, cachon@wharton.upenn.edu

Fuqiang Zhang
Paul Merage School of Business, University of California, Irvine, California 92697-3125, fzhang@uci.edu

Any buyer that depends on suppliers for the delivery of a service or the production of a make-to-order component should pay close attention to the suppliers' service or delivery lead times. This paper studies a queueing model in which two strategic servers choose their capacities/processing rates and faster service is costly. The buyer allocates demand to the servers based on their performance; the faster a server works, the more demand the server is allocated. The buyer's objective is to minimize the average lead time received from the servers. There are two important attributes to consider in the design of an allocation policy: the degree to which the allocation policy effectively utilizes the servers' capacities and the strength of the incentives the allocation policy provides for the servers to work quickly. Previous research suggests that there exists a trade-off between efficiency and incentives, i.e., in the choice between two allocation policies a buyer may prefer the less efficient one because it provides stronger incentives. We find considerable variation in the performance of allocation policies: Some intuitively reasonable policies generate essentially no competition among servers to work quickly, whereas others generate too much competition, thereby causing some servers to refuse to work with the buyer. Nevertheless, the trade-off between efficiency and incentives need not exist: It is possible to design an allocation policy that is efficient and also induces the servers to work quickly. We conclude that performance-based allocation can be an effective procurement strategy for a buyer as long as the buyer explicitly accounts for the servers' strategic behavior.

Key words: game theory; joining behavior; Nash equilibrium; procurement; sourcing; supplier management
History: Accepted by Wallace J. Hopp, stochastic models and simulation; received February 1, 2005. This paper was with the authors 8 months for 2 revisions.

Fast service is clearly important. Less obvious is how to go about obtaining fast service from suppliers or service providers. One technique is to make servers compete by allocating business to them based on their performance, i.e., the faster server is rewarded with a greater share of demand. For example, Sun Microsystems maintains multiple memory chip suppliers and allocates demand with a scorecard system. A score that depends on a number of factors, delivery lead time among them is periodically assigned to each supplier, and a supplier's allocation of Sun's business increases as they improve their score relative to the other suppliers (Farlow et al. 1996). GE Lighting and Air Products and Chemicals also allocate demand towards better-performing suppliers (Pyke and Johnson 2003).

This paper studies, in the context of a stylized queueing model, the issue of how performance-based demand allocation can induce competition among suppliers to obtain faster service or delivery lead times. A precursor to this line of research is the extensive body of work on queue-joining behavior, pioneered by Naor (1969). That literature focuses on the behavior of strategic customers/jobs: e.g., whether or not to join a queue (e.g., Naor 1969), or which of several queues to join (Bell and Stidham 1983). It is generally found that the behavior of individual jobs creates externalities on other jobs (e.g., overcongestion of the faster server). (See Hassin and Haviv 2003 for a review of the queue-joining literature with strategic customers/jobs.) Those externalities do not occur in our setting because a single buyer controls all of the jobs. Instead, we have strategic servers—servers that can regulate how fast they work, and working faster is costly.

In our model the buyer pays a fixed amount for each job, so the buyer's task is to choose an allocation policy to minimize the average lead time to complete jobs. We study allocation policies that can be classified into two groups, state-dependent policies (the

allocation of a job to a server depends on the servers' current workload) and state-independent policies (the allocation of a job does not depend on the number of jobs currently in the servers' queues).

With nonstrategic servers it is clear that a state-dependent policy can deliver faster lead times than a state-independent policy because, in part, a state-independent policy risks allocating jobs to busy servers while other servers remain idle, i.e., a state-dependent policy can do a better job of pooling the servers' capacities.[1] However, are state-dependent policies better when the servers are strategic? Suppose a state-independent policy induces servers to work more quickly than a state-dependent policy. Then the buyer may be better off with a state-independent policy even though the system's capacity is not as effectively utilized. In other words, incentives may trump efficiency. In fact, Gilbert and Weng (1998) arrive at that conclusion. Nevertheless, there are several reasons why this might not be the best conclusion. First, we show that there is an error in their equilibrium existence proof, so it is not always meaningful to compare their two allocation policies. Second, and more importantly, they do not compare optimal policies. We compare the buyer's best state-dependent policy with the buyer's best state-independent policy and find that the buyer is better off with the state-dependent policy, i.e., the buyer can have both incentives and efficiency. In general, we find that there is considerable variation in the performance of intuitively reasonable policies. For example, the buyer's optimal state-independent policy with nonstrategic servers is found to perform poorly in the presence of strategic servers, and proportional allocation, which is probably the most intuitive allocation policy, can be the worst performer of the policies we consider.

The next section describes our model in detail. Section 2 expands upon the related literature. Section 3 studies the buyer's allocation policy choice and the competition between servers under several different allocation policies. Section 4 discusses several extensions to the model. The final section concludes with a summary of our results.

1. The Model

A buyer procures a good (e.g., a make-to-order component, as in the Sun Microsystems example) or a service. For ease of exposition, we assume a service is procured. There are two servers. (Most of our results extend to more than two servers; see Zhang 2004 for details.) Demand for the service arrives according to a Poisson process with rate λ. Each demand is referred to as a job and all jobs are eventually completed. Server i's average service rate is μ_i and service times are exponentially distributed. We refer to μ_i as server i's capacity and $\mu = (\mu_1, \mu_2)$ denotes the capacity vector. A server with capacity μ_i incurs a capacity cost at rate $c(\mu_i)$, no matter whether the capacity is utilized or idle, where $c(0) = 0$, $c'(\cdot) > 0$, and $c''(\cdot) \geq 0$ are assumed. The servers' variable cost per job is normalized to zero.

We say that a job is allocated to a server when it is certain that server will complete the job. The buyer pays R per allocated job. We assume $R > r_1$, where

$$r_1 = c(\lambda/2)/(\lambda/2),$$

because it is the minimal requirement for the suppliers to earn a nonnegative profit and deliver finite lead times (see Zhang 2004). We assume R is exogenous: There could be an industry standard price that the buyer is unable to negotiate away from, or the price could be set via negotiations that involve issues beyond the scope of this model.

The buyer controls her allocation policy (i.e., how jobs are allocated to servers) and the servers choose their capacities. The buyer seeks to minimize the average delivery lead time over an infinite-horizon subject to the constraint that each server earns a nonnegative profit, and the servers seek to maximize their average profit:

$$\pi_i(\mu) = R\lambda_i(\mu) - c(\mu_i), \tag{1}$$

where $\lambda_i(\mu)$ is the rate at which server i is allocated jobs.[2] Hence, we assume that the buyer and the servers do not discount future cash flows and that they expect a long-term relationship. We focus on equilibria in which the servers adopt open-loop strategies, i.e., strategies that are independent of the history of play. As a result, this infinite-horizon capacity game among servers can be analyzed as a single-decision capacity game. Previous research on strategic servers also restricts attention to open-loop strategies. In §3.3 we discuss lead time-based allocation rather than capacity-based allocation.

2. Literature Review

Kalai et al. (1992) were the first to study strategic servers, but they only consider a simple state-dependent policy in which jobs are allocated to idle

[1] Pooling is not necessarily a good idea if servers have significantly different capacities. Rubinovitch (1985a) characterizes the conditions under which a job should never be allocated to the slow server in a two-server queueing system.

[2] Note that servers are paid for allocated jobs rather than completed jobs. If they were paid for completed jobs then their profit function would be $\pi_i(\mu) = R\min\{\mu_i, \lambda_i(\mu)\} - c(\mu_i)$. The equilibrium analysis of this profit function is significantly more complex due to the kink created by the min function. Nevertheless, our qualitative results are not different. See Zhang (2004) for details.

servers with equal probability. Gilbert and Weng (1998) expand upon their model to include a state-independent allocation policy that allocates jobs to servers immediately upon arrival. They conclude that a state-independent policy can be better for the buyer than a state-dependent policy. Our results are different, as we explain in detail in the subsequent sections. Christ and Avi-Itzhak (2002) extend those models to include customer balking, but we do not have balking.

Ha et al. (2003) study the competition between two suppliers serving one buyer, in which delivery frequency is an element of the buyer's allocation decision. However, they study deterministic demand, so although they consider issues similar to ours, a direct comparison between their work and ours is not meaningful.

There are papers that compare sole sourcing versus dual sourcing, whereas we assume that a dual-sourcing strategy has been adopted: e.g., Anton and Yao (1989, 1992), Anupindi and Akella (1993), Benjaafar et al. (2007), Seshadri (1995), and Seshadri et al. (1991). See Minner (2003) and Elmaghraby (2000) for reviews of the literature on sourcing strategies.

There are papers that study a buyer's procurement policy when there are multiple suppliers with exogenously determined characteristics: e.g., Bonser and Wu (2001), Chen et al. (2001), Li and Kouvelis (1999), Martinez de Albeniz and Simchi-Levi (2003), Sedarage et al. (1999), and Talluri (2002). In our model the servers' lead times depend on their choices and the buyer's allocation policy.

Several papers study coordination and competition in supply chains with multiple suppliers: Bernstein and DeCroix (2004); Wang and Gerchak (2003); and Nagarajan and Bassok (2003). In these papers, limited capacity leads to demand truncation rather than slower delivery times. Bernstein and de Vericourt (2005) consider a market with multiple suppliers and multiple buyers. Their suppliers have fixed processing rates and compete by offering different lead times to buyers, which they obtain via holding inventory.

There are a number of papers that study server competition in which firms choose operational strategies to adjust their delivery times: e.g., Allon and Federgruen (2003), Cachon and Harker (2002), Chayet and Hopp (2002), Lederer and Li (1997), and So (2000). In those papers the structure of how firms compete is exogenous, whereas in our model it is determined by the buyer via her allocation policy.

There is literature on capacity allocation (e.g., Cachon and Lariviere 1999a, b, c; Deshpande and Schwarz 2002), in which a single manufacturer allocates scarce capacity among multiple buyers. Although allocation policies similar to ours are implemented, those models are analytically quite different.

Li (1992) and Armony and Plambeck (2005) study models in which a buyer submits duplicate orders to multiple suppliers. In our model, each job is allocated to a single server, but we briefly discuss order duplication in §4.

3. Allocation and the Servers' Capacity Game

Our model can be analyzed in two interdependent parts. The first part is the buyer's allocation policy choice—i.e., how will the buyer allocate jobs among the two servers'. The second part is the capacity choice game played between the servers, which clearly depends on the particular allocation policy the buyer has selected. Furthermore, the attractiveness of an allocation policy to the buyer depends on the capacities chosen by the servers, as well as how jobs are routed through the system. We treat these two parts sequentially.

The set of allocation policies can be divided into two broad classes: state-independent policies and state-dependent policies. With a state-independent policy, the buyer allocates jobs to servers based only on their capacities (which are inferred from past allocations and resulting delivery times) and not on the current state of the system (e.g., how many jobs are allocated to each server, which server is idle, etc.). Because no current information is utilized with a state-independent policy, the buyer immediately allocates a job to a server upon its arrival, i.e., there is no benefit in waiting to allocate a job if waiting does not change the allocation decision process. In contrast, with a state-dependent allocation policy the buyer allocates jobs based on the current state of the system. For example, the buyer may choose to allocate jobs only to idle servers.

Given a fixed-capacity vector, the buyer's optimal state-dependent policy is clearly never worse (and can be strictly better) than the buyer's optimal state-independent policy because state-independent policies are a subset of the set of state-dependent policies. To be more specific, assume both servers choose capacity μ_i so that it is optimal for the buyer to allocate half of the jobs to each server. The optimal state-dependent policy allocates jobs only to idle servers, and so the average lead time, $W_{sd}(\mu_i)$, is equivalent to an $M/M/2$ queueing system,

$$W_{sd}(\mu_i) = \frac{\mu_i}{\mu_i^2 - (\lambda/2)^2}.$$

The optimal state-independent policy allocates jobs upon arrival to servers with equal probability, which yields an average lead time, $W_{si}(\mu_i)$, that is equivalent to two $M/M/1$ systems,

$$W_{si}(\mu_i) = \frac{1}{\mu_i - \lambda/2}.$$

Assuming stable systems, $\mu_i > \lambda/2$, it is intuitive that the state-dependent lead time is faster than the state-independent lead time, $W_{sd}(\mu_i) < W_{si}(\mu_i)$, because the state-dependent policy does a better job of pooling the servers' capacities. With the state-dependent policy a job is never waiting while there is an idle server, but that inefficient outcome can occur with the state-independent policy.

In addition to how jobs are routed through the system, the buyer's lead time depends on the capacities chosen by the servers. Again assuming that the servers choose identical capacities, it is easy to see that both $W_{sd}(\mu_i)$ and $W_{si}(\mu_i)$ are decreasing in μ_i, i.e., the buyer's lead time with either type of allocation is reduced as the servers work faster. Because working faster is costly to the servers, there exists a maximum rate, $\bar{\mu}$, at which the servers earn zero profit given that they are allocated half of the jobs, i.e., $\bar{\mu}$ is the solution to $c(\bar{\mu}) = R\lambda/2$. From the buyer's perspective, the ideal state-dependent allocation policy induces the servers to choose capacity $\bar{\mu}$ and routes jobs so that the resulting lead time is $W_{sd}(\bar{\mu})$. Similarly, the ideal state-independent policy induces the servers to choose capacity $\bar{\mu}$ and routes jobs so that the lead time is $W_{si}(\bar{\mu})$.[3] It remains to be determined whether those ideals can be achieved, i.e., does there exist an allocation policy that achieves $\bar{\mu}$ as an equilibrium outcome of the servers' capacity game? If so, then clearly the optimal state-dependent policy would be strictly better for the buyer than the optimal state-independent policy.

3.1. State-Independent Allocation Policies

Bell and Stidham (1983) identify the state-independent allocation policy, which we call Bell-Stidham allocation, that minimizes the buyer's lead time for any fixed-capacity vector, μ:

$$\lambda_i(\mu) = \begin{cases} \mu_i - \left(\mu_i^{1/2} \Big/ \sum_{j=1}^{\hat{n}} \mu_j^{1/2}\right)\left(\sum_{j=1}^{\hat{n}} \mu_j - \lambda\right) & \text{for } i \le \hat{n}, \\ 0 & \text{for } i > \hat{n}, \end{cases} \tag{2}$$

where the servers' capacities are sorted in decreasing order and $\hat{n} \le 2$ is the largest integer, such that $\lambda_{\hat{n}}(\mu) \ge 0$.[4] This allocation rule equates the marginal change in the average number of jobs at each queue with respect to the arrival rate. Naturally, Bell-Stidham allocation assigns half of the jobs to each server when the servers have the same capacity, μ_{bs}, thereby achieving the lead time $W_{si}(\mu_{bs})$.

Bell-Stidham allocation was designed for nonstrategic servers. With strategic servers, according to Theorem 1, a symmetric equilibrium exists in this capacity game only under certain conditions. The capacity cost function restriction is relatively mild, but the restrictions on R are significant. (All proofs are in the appendix.)

THEOREM 1. *With Bell-Stidham allocation,* (2), *if* $R > r_2 = 2c'(\lambda/2)$, $c'''(\mu_i) \ge 0$, *and* $\pi_i(\mu_{bs}) \ge 0$, *where* μ_{bs} *is the unique solution to*

$$\left(\frac{R}{4}\right)\left(1 + \frac{\lambda/2}{\mu_{bs}}\right) - c'(\mu_{bs}) = 0, \tag{3}$$

then $\mu_i = \mu_{bs} > \lambda/2$ *is the unique symmetric Nash equilibrium.*

An equilibrium (with finite lead times) may fail to exist with Bell-Stidham allocation because the buyer's price may be too low, $R \le r_2$: The servers do not feel the need to build enough capacity to provide a stable system (i.e., they prefer to work at 100% utilization than to compete for additional demand by working more quickly and operating at less than 100% utilization). (Note that because $c(\cdot)$ is convex, it is straightforward to show that $r_1 < r_2$.) Alternatively, an equilibrium may fail to exist because the buyer pays too much, thereby causing so much competition between the servers that they both cannot earn a positive profit.[5] Furthermore, it is apparent from (3) that the servers may not choose in equilibrium the buyer's ideal capacity, i.e., $\mu_{bs} \ne \bar{\mu}$ is possible.

Although Bell-Stidham allocation is optimal for the buyer for any given capacity vector, it does not take into consideration the behavior of strategic servers, and, as a result, it does not necessarily provide the correct incentives for servers to choose a desirable capacity vector. With strategic servers it is important to recognize that the buyer's allocation policy need not be optimal for all capacity vectors (as is Bell-Stidham). The role of the allocation policy is to establish incentives for the servers to converge

[3] We assume the buyer desires to have two symmetric servers. Given that the servers have the same capacity cost function, it is either optimal for the system to have one server that is allocated all jobs or two servers that are allocated half of the jobs, where the latter is more likely as the capacity cost function becomes more convex. There could be other reasons for maintaining multiple servers even if the capacity cost function suggests one server would be optimal. We do not attempt to model those alternative reasons, so we assume throughout that the buyer desires to dual source and equally divide jobs between the servers.

[4] They also provide results for $M/G/1$ queues and allow waiting time costs to vary across queues. In this application the waiting time cost is naturally the same across all queues. We discuss in §4 our results with nonexponential service times.

[5] For example, with a quadratic capacity cost function it can be shown that there exists an upper bound, r_3, such that there does not exist an equilibrium with $R > r_3$.

to a particular capacity equilibrium that is desirable for the buyer, ideally $(\bar{\mu}, \bar{\mu})$. As a result, it is worthwhile to consider other allocation policies that achieve an equal division of jobs in equilibrium, as with Bell-Stidham, but allocate jobs differently than Bell-Stidham for nonequilibrium/nonsymmetric capacities.

Gilbert and Weng (1998) propose balanced allocation: With balanced allocation the buyer attempts to equalize (i.e., balance) the servers' lead times for all capacity vectors (and only fails to do so if all jobs are allocated to one server because of a large disparity in their processing rates):

$$\lambda_i(\mu) = \begin{cases} \lambda & \lambda + \mu_j \le \mu_i \\ \left(\mu_i - \frac{1}{2}(\mu_i + \mu_j - \lambda)\right)^+ & \text{otherwise.} \end{cases} \tag{4}$$

THEOREM 2. *With balanced allocation, if* $R \ge r_2 = 2c'(\lambda/2)$, $c''(\mu_i) > 0$, *and* $c'(\mu_b) \ge c(\mu_b)/\lambda$, *where* μ_b *is the unique solution to* $c'(\mu_b) = R/2$, *then* $\{\mu_b, \mu_b\}$ *is the unique Nash equilibrium and the servers' average lead times are finite. Otherwise, there does not exist an equilibrium with finite lead times.*

As with Bell-Stidham allocation, balanced allocation leads to a symmetric equilibrium, but the two allocation policies need not result in the same capacity, $\mu_{bs} \ne \mu_b$, and balanced allocation also generally results in less than the buyer's desired capacity, $\mu_b \le \bar{\mu}$. Furthermore, three conditions are needed for an equilibrium to exist with balanced allocation. First, balanced allocation requires that the buyer's price is sufficiently high, $R \ge r_2$, otherwise the reward for working fast is insufficient to provide an incentive to work. Second, the capacity cost function must be strictly convex, $c''(\mu_i) > 0$, which rules out the important case of linear capacity costs. Gilbert and Weng (1998) correctly recognized those first two conditions, but did not recognize the necessary third condition, $c'(\mu_b) \ge c(\mu_b)/\lambda$, which requires the servers to earn a nonnegative profit in equilibrium (e.g., with a quadratic cost function $c(\mu_i) = a\mu_i^2 + b\mu_i$, $a > 0$, this condition translates into $R \le 2(2a\lambda + \sqrt{b^2 + 4a^2\lambda^2})$). They erred by believing that each server's profit function is globally concave. In fact, it is concave and decreasing for $\mu_i \in [0, \mu_j - \lambda]$ and concave for $\mu_i > \mu_j - \lambda$. Hence, each server's global optimum is either the maximum of the first concave range, $\mu_i = 0$, or the maximum of the second concave range, $\mu_i > \mu_j - \lambda$. As a result, each server's reaction function (the optimal capacity given the capacity of the other server) harbors a discontinuity, which creates the possibility of no equilibrium. However, if an equilibrium exists, then Gilbert and Weng (1998) correctly identify it.

An alternative allocation policy is needed that can be parameterized so as to adjust up or down, as needed, the level of competition between the servers. We offer two such policies: linear allocation and proportional allocation. With linear allocation,

$$\lambda_i(\mu) = \begin{cases} \theta\mu_i^\rho - \frac{1}{\hat{n}}\left(\theta \sum_{j=1}^{\hat{n}} \mu_j^\rho - \lambda\right) & \text{for } i \le \hat{n} \\ 0 & \text{for } i > \hat{n}, \end{cases} \tag{5}$$

where the servers' capacities are sorted in decreasing order, $\theta > 0$, $0 < \rho \le 1$, and $\hat{n} \le 2$ is the largest integer such that $\lambda_{\hat{n}} \ge 0$ and $\mu_{\hat{n}} > 0$. A server does not necessarily receive a positive allocation even if the server builds some capacity, but a server surely receives no allocation if the server builds no capacity. If $\theta = 1$ and $\rho = 1$, then linear allocation is almost identical to balanced allocation: The only exception is the additional $\mu_{\hat{n}} > 0$ requirement to receive a positive allocation. (That reasonable requirement facilitates the uniqueness equilibrium proof.) Hence, linear allocation can be considered a generalization of balanced allocation.

The parameters θ and ρ could potentially enable linear allocation to achieve many different capacity vectors as an equilibrium to the servers' capacity game. However, as already discussed, the buyer's desired outcome from the servers' capacity game is $(\bar{\mu}, \bar{\mu})$ with an even division of jobs between the servers. According to the next theorem, linear allocation can achieve that objective. Hence, linear allocation is an optimal state-independent allocation policy.

THEOREM 3. *Given linear allocation*:

(i) *If* $c''(\mu_i) > 0$, $\theta = 2c'(\mu_l)/R$, *and* $\rho = 1$, *then* $\mu_i = \mu_l = \bar{\mu}$ *for all* i *is a unique Nash equilibrium and the average lead times are finite.*

(ii) *If* $c(\mu_i) = b\mu_i$ $(b > 0)$, $\theta = 4\mu_l^{1/2}c'(\mu_l)/R$, *and* $\rho = 1/2$, *then* $\mu_i = \mu_l = \bar{\mu}$ *for all* i *is the unique Nash equilibrium and the average lead times are finite.*

The parameters provided in Theorem 3 are not the only ones that achieve our objective (that $(\bar{\mu}, \bar{\mu})$ is the unique Nash equilibrium), so we choose intuitive values for ρ: With strictly convex capacity cost the ρ parameter is not necessary (hence, set to $\rho = 1$), but with a linear capacity cost $\rho < 1$ is necessary to create an interior optimum for each server.

Proportional allocation is another policy that can be parameterized to adjust the level of competition between the servers. With proportional allocation, server i's share of the buyer's jobs is

$$\lambda_i(\mu) = \left(\frac{\mu_i^\beta}{\mu_1^\beta + \mu_2^\beta}\right)\lambda, \tag{6}$$

where $\beta \ge 1$ is a parameter. In particular, increasing β raises the intensity of competition, thereby allowing the buyer to achieve the desired capacity vector,

$(\bar{\mu}, \bar{\mu})$. Hence, proportional allocation can also be an optimal state-independent allocation policy. However, because the servers' profit functions are not necessarily well behaved as β is increased, Theorem 4 provides results only for a quadratic capacity cost function.

THEOREM 4. *Given proportional allocation and a quadratic capacity cost function* $c(\mu_i) = a\mu_i^2 + b\mu_i$, $a \geq 0$, $b \geq 0$, $a + b > 0$, *if*

$$\beta = \frac{2\bar{\mu}c'(\bar{\mu})}{c(\bar{\mu})},$$

where $c(\bar{\mu}) = R\lambda/2$ (*i.e.*, $\bar{\mu}$ *is the server's break-even capacity*), *and* $R > r_1 = c(\lambda/2)/(\lambda/2)$, *then* $\mu_i = \bar{\mu}$ *for all* i *is the unique Nash equilibrium and average lead times are finite*.

Although $\beta > 1$ is desirable for the buyer, it is worthwhile to mention that $\beta = 1$ yields an intuitively appealing allocation mechanism: With $\beta = 1$ a server's demand share equals the server's share of total capacity and the servers' utilizations are equated (i.e., each server has the same number of jobs on average). Recall that Bell-Stidham allocation equates the marginal change in the number of jobs at each server with respect to that server's arrival rate. However, existence of an equilibrium with $\beta = 1$ requires the buyer to pay a sufficiently large price and the servers' capacities are less than ideal for the buyer, $\mu_p < \bar{\mu}$.

THEOREM 5. *With proportional allocation and* $\beta = 1$, *if* $R > r_2 = 2c'(\lambda/2)$, *then* $\mu_i = \mu_p$ *for all* i *is a unique Nash equilibrium with finite lead times*, *where* μ_p *is the unique solution to*

$$c'(\mu_p) = \left(\frac{R}{4}\right)\left(\frac{\lambda}{\mu_p}\right).$$

Otherwise, a Nash equilibrium does not exist with finite lead times.

3.2. State-Dependent Allocation

The simplest state-dependent policy is common-queue allocation, first studied by Kalai et al. (1992): Jobs are only allocated to idle servers, where each idle server is equally likely to be allocated a job, and jobs are maintained on a queue if both servers are occupied. For convenience, the following lemma repeats their results.

LEMMA 6. *Given that* $c''(\mu_i) > 0$ *and the buyer implements common-queue allocation, let* μ_c *be the unique solution to*

$$c'(\mu_c) = \frac{R\lambda^2}{2\mu_c(2\mu_c + \lambda)}.$$

If $R > r_2 = 2c'(\lambda/2)$, *then* $\{\mu_c, \mu_c\}$ *is the unique Nash equilibrium in the capacity game and the servers' average lead times are finite. If* $R \leq r_2$, *then there does not exist an equilibrium with finite lead times*.

Common-queue allocation has the desirable feature that it pools the capacities of the servers (there are never waiting jobs and idle servers at the same time). Hence, with nonstrategic and identical servers, common queue is in fact optimal for the buyer. However, an equilibrium with finite lead times does not exist with common-queue allocation if the price is too low, $R \leq r_2$. Furthermore, Gilbert and Weng (1998) demonstrate that with strategic servers common queue can be worse for the buyer than balanced allocation because it does not provide sufficient incentives for the servers to work quickly. Hence, a state-dependent allocation policy may actually perform worse than a state-independent policy.

Although common queue is optimal for the buyer given symmetric capacities, it is not optimal for the buyer with asymmetric capacities. Intuitively, if one server is much slower than the other server, then the buyer may be better off allocating a job to the busy fast server than to the idle slow server; e.g., a fast server may be able to complete two jobs faster than the slow server can complete one job. This intuition suggests a threshold allocation policy that is implemented as follows. One server is labeled the primary server and the other the secondary server. A single parameter, $m \in \{0, 1, 2, \ldots\}$, regulates how jobs are allocated to the primary and secondary servers: allocate a job to the primary server if the primary server is idle or if the primary server has fewer than m jobs in queue; allocate a job to the secondary server only if the secondary server is idle, the primary server is busy, and has m jobs in queue. It is natural to think of the faster server as the primary server, but the policy can also be implemented with the slower server designated as the primary.

Given nonstrategic servers, Rubinovitch (1985b) provides a numerical method to evaluate the system's performance under threshold allocation, and Lin and Kumar (1984) prove that a threshold policy is the buyer's optimal allocation with two servers, i.e., the average time in the system for each unit is minimized. Additional proofs are available from Koole (1995) and Walgrand (1984).

It is intuitive that as the threshold parameter, m, increases, the primary server's share of the buyer's demand increases and the secondary server's share decreases. With $m = \infty$, the primary server earns the buyer's entire demand, while the secondary server is never allocated a job. Hence, by varying which server is designated the primary and by randomizing between different m values, the buyer is able to allocate to the faster server any portion of the buyer's demand.[6] As a result, it is possible to design a threshold policy in which server i's allocation exactly equals

[6] Even with $m = 0$, the faster server, when designated the primary, can earn more than 50% of the buyer's demand. Threshold

his allocation with linear allocation for any chosen capacities. Servers only care about their share of the buyer's jobs, not how that allocation is achieved or the resulting lead time for the buyer. Therefore, if the described threshold policy is used, the equilibrium in the capacity game is equivalent to the equilibrium with linear allocation. Furthermore, in equilibrium the servers have equal capacity, so the threshold is $m = 0$, i.e., in equilibrium the servers build capacity as if linear allocation were implemented, but the system actually achieves the same lead time as common-queue allocation. Although the techniques in Rubinovitch (1985b) allow for the evaluation of the proper thresholds, a threshold policy is clearly not as simple to evaluate as the other allocation policies we discuss. However, in theory, it provides in equilibrium the maximum capacity like linear allocation, while also providing the operational efficiency of common-queue allocation. Hence, it is an optimal state-dependent policy for the buyer. We conclude that there need not exist a trade-off between incentives and efficiency: The optimal state-dependent policy, threshold allocation, performs better than the optimal state-independent policy, linear allocation.

Additional comparisons among the policies can be made via some graphical examples. For each allocation policy, Figures 1 and 2 show the relationship between R and the equilibrium lead times with two examples: $c(\mu_i) = 4\mu_i$ and $\lambda = 1$; and $c(\mu_i) = 4\mu_i^2$ and $\lambda = 1$. We see from these figures that for a given price the buyer's lead time can vary considerably. In all cases, common-queue allocation and proportional allocation with $\beta = 1$ perform poorly. Bell-Stidham allocation gives intermediate performance. Balanced allocation performs reasonably well when an equilibrium exists, but an equilibrium exists for a relatively limited range of prices (it never exists with linear capacity cost). Overall, threshold allocation is clearly the best, but linear allocation, especially given its simplicity, is a good second choice.

The next lemma further explores the difference between linear and threshold allocation.

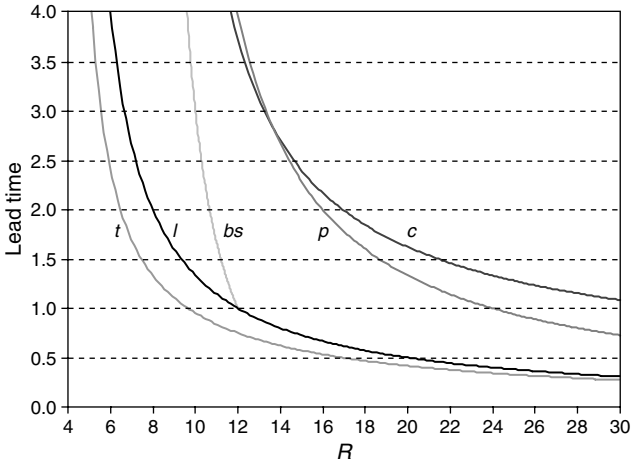
Lemma 7. *Define*

$$z(R) = W_{sd}(\mu_t(R))/W_{si}(\mu_l(R)),$$

where $\mu_t(R)$ and $\mu_l(R)$ are the equilibrium capacities under threshold and linear allocations, respectively, when the price is R. Recall that $\mu_t(R) = \mu_l(R)$, i.e., for a fixed wholesale price, threshold and linear allocations generate the same capacity. The ratio $z(R)$ is concave and increasing from $1/2$ to 1.

allocation can assign less demand to the faster server only if the faster server is designated the secondary server.

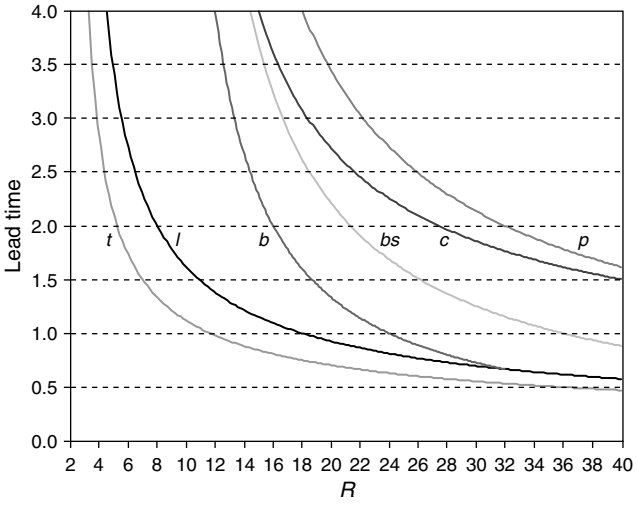
Figure 1 **The Lead Time Received by the Buyer as a Function of the Price Paid, R, and the Allocation Policy with Capacity Cost $c(\mu_i) = 4\mu_i$ and $\lambda = 1$**



Note. t = threshold policy, l = linear allocation, bs = Bell-Stidham, c = common queue, p = proportional allocation with $\beta = 1$. Balanced allocation is not included because an equilibrium does not exist in this setting.

The comparison between threshold and linear allocation is intuitive: If system utilization is quite high because R is low, then threshold allocation has a single queue with a large number of jobs, whereas linear allocation has two queues with a large number of jobs (i.e., threshold's lead time is half of linear's lead time). However, if system utilization is quite low because R is high, then jobs never wait with either allocation policy. Although Lemma 7 indicates that linear allocation is significantly worse than threshold allocation when the buyer's price is low, this result is somewhat mis-

Figure 2 **The Lead Time Received by the Buyer as a Function of the Price Paid, R, and the Allocation Policy with Capacity Cost $c(\mu_i) = 4\mu_i^2$ and $\lambda = 1$**



Note. t = threshold policy, l = linear allocation, b = balanced allocation, bs = Bell-Stidham, c = common queue, p = proportional allocation with $\beta = 1$.

leading. Now suppose that the buyer is able to modify her price somewhat. Let R_t be the price with threshold allocation and let R_l be the price with linear allocation and choose these prices such that they lead to the same delivery lead time, $W_{sd}(\mu_t(R_t)) = W_{si}(\mu_l(R_l))$. According to the next lemma, if R_t is either low or high, then there is a small price premium needed with linear allocation to achieve the same lead time.

Lemma 8. *Let ρ be the system's utilization in equilibrium. $R_l/R_t \to 1$ as either $\rho \to 1$ or $\rho \to 0$.*

3.3. Lead Time-Based Allocation

This section considers whether the buyer could do better (or at least as well) with an allocation policy based on the servers' lead times rather than based on their capacities. In a lead time-based allocation, the buyer announces the demand share function λ_i in terms of servers' lead time vector $W = (W_1, W_2) > 0$, the servers submit their bids on lead times, demand shares are determined, and each server i builds capacity $\mu_i(W_i, \lambda_i(W))$ to fulfill its lead-time bid, where μ_i is a decreasing function of W_i. Assume

$$\lambda_1(W) - \lambda_1(W_\varepsilon) < \mu_1(W_1, \lambda_1(W)) - \mu_1(W_1 + \varepsilon, \lambda_1(W_\varepsilon)) \quad (7)$$

where $W_\varepsilon = (W_1 + \varepsilon, W_2)$ and $\varepsilon > 0$: if server 1 promises a longer lead time, then server 1's required capacity to achieve that lead time decreases faster than server 1's demand allocation. The analogous assumption is taken for the other server as well. This assumption holds, for example, when each server operates an $M/M/1$ queue, in which case

$$\mu_i(W_i, \lambda_i(W)) = 1/W_i + \lambda_i(W). \quad (8)$$

Lead time-based allocation is analytically cleaner than capacity-based allocation because there is no issue with the stability of the queues: By definition, the buyer's lead time is positive and finite for any strategic choice vector of the servers, whereas with capacity-based allocation the servers may fail to choose a sufficient capacity to yield a finite lead time for the buyer. However, according to the next lemma, analytical tractability can come with a price.

Lemma 9. *Consider any continuous lead-time allocation with λ_i decreasing in W_i. If (W_1, W_2) is a Nash equilibrium with corresponding demand shares (λ_1, λ_2) and $\mu_i(W_i, \lambda_i(W))$ satisfies (7), then $\mu_i(W_i, \lambda_i(W)) \le \hat{\mu}$ for all i, where $\hat{\mu}$ is the solution to $c'(\hat{\mu}) = R$. (If $c(\mu)$ is linear, let $\hat{\mu} = \infty$.)*

Recall that $\bar{\mu}$ is the servers' maximum capacity (i.e., $c(\bar{\mu}) = R\lambda/2$) and the capacity achieved with linear or threshold allocation based on capacities. It is possible that the maximum achievable capacity with a lead time-based allocation policy, $\hat{\mu}$, is less than the maximum with a capacity-based allocation policy, $\bar{\mu}$. We demonstrate this with two examples in which the relationship between a server's lead time and its capacity is given by (8), i.e., a state-independent allocation policy is implemented. First, suppose the capacity cost function is quadratic, $c(\mu) = a\mu^2 + b\mu$ and $a > 0$. Then $\bar{\mu} > \hat{\mu}$ for all $R \in (r_1, a\lambda + \sqrt{b^2 + a^2\lambda})$, i.e., for sufficiently small R in the feasible range ($R > r_1$) the buyer cannot design a continuous allocation policy based on the servers' lead times that achieves the maximum capacity, $\bar{\mu}$. Next suppose $c(\mu) = a\mu^\gamma$ for $a > 0$ and $\gamma > 1$. In this case,

$$\frac{\bar{\mu}}{\hat{\mu}} = \frac{(R\lambda/2a)^{1/\gamma}}{(R/\gamma a)^{1/(\gamma-1)}} = \left(\gamma\left(\frac{r_1}{R}\right)^{1/\gamma}\right)^{1/(\gamma-1)},$$

where recall that $r_1 = c(\lambda/2)/(\lambda/2) = a(\lambda/2)^{\gamma-1}$. It follows that $\bar{\mu} > \hat{\mu}$ when $R \in (r_1, r_1\gamma^\gamma)$, i.e., lead time-based allocation is likely to be inferior to capacity-based allocation because the buyer's price is low and as the capacity cost function becomes more convex (γ increases).

Despite the one-to-one relationship between a server's lead time and the server's capacity for a fixed allocation, lead time-based allocation may not be as effective as capacity-based allocation because lead time-based allocation has a self-restraining property that dampens competition among the servers: Committing to a higher service level requires more investment than committing to a higher capacity. A similar result is obtained in Cachon and Zipkin (1999) in the context of inventory management in a serial supply chain with two independent firms: With nonstrategic firms the optimal policy can be implemented as either a set of installation base-stock policies or as a set of echelon base-stock policies (briefly, these policies differ in what information they use), but with strategic firms these two approaches yield different equilibrium results.

4. Discussion

This section discusses several modeling issues. Although we assume exponential processing times, some of our results extend to more general processing time distributions. As in Bell and Stidham (1983), suppose μ is the service rate and the service time S has first moment $E(S) = 1/\mu$ and second moment $E(S^2) = b\mu^{-2}$. The variance is then $(b-1)\mu^{-2}$ and the coefficient of variation is constant, $(b-1)^{1/2}$. For an $M/G/1$ queue with the above service time distribution, the average lead time (duration in the system) is

$$W(\lambda) = \frac{1}{\mu} + \frac{b\lambda}{2\mu(\mu - \lambda)}.$$

Balanced, linear, proportional, Bell-Stidham, and threshold allocations readily extend to this general distribution because demand is allocated based only on the servers' capacities and not on their lead times. However, the extension is not straightforward for common queue because then the servers' shares of demand are endogenously determined.

Throughout our analysis we have assumed that each job is processed by only one server. In practice, there are examples in which firms duplicate their orders across multiple suppliers or servers (see Armony and Plambeck 2005, Li 1992, Yoffie 1990). If order duplication is feasible, then it is ideal from the point of view of system efficiency: Even if there is only one job in the system all servers are working at their full rate. However, as we have demonstrated, it is also important for an allocation policy to provide sufficient incentives for strategic servers to work hard. Zhang (2004) demonstrates that order duplication performs poorly on incentives, so poorly that its overall performance tends to be worse than linear allocation. Hence, even if operating conditions are ideal for order duplication, a buyer should avoid order duplication.

We use demand allocation as the motivator to provide fast service, but other motivators may exist. For example, if the buyer has some control over the price, R, then raising the price, as we see in Figures 1 and 2, generates faster service (but with some allocation policies it also eliminates the existence of an equilibrium). The buyer could make a trade-off between the higher price paid and the faster service received. Nevertheless, unless the price paid is extremely high, there remains considerable variation in the performance of the allocation policies.

Instead of allocating demand, the buyer could try to motivate faster service by posting a payment schedule that is contingent on the servers' capacities or lead times. For example, suppose the buyer wants each server to build $\mu_i^* > \lambda/2$ capacity. This is achievable with the following price schedule, $R(\mu_i)$, $R''(\mu_i) < 0$, $R'(\mu_i^*)(\lambda/2) = c'(\mu_i^*)$, and $R(\mu_i^*)(\lambda/2) = c(\mu_i^*)$: The first condition ensures a unique μ_i maximizes the server's profit, the second condition ensures that μ_i^* is optimal for the server, and the third condition makes the server's profit condition binding. It is also possible that the $R(\mu_i)$ schedule could be implemented with a fixed price and late fees, because then the late fees paid are contingent on the chosen capacity. (See Cachon and Zhang 2006 for a similar model with sole sourcing and late fees.) Our model does not address whether demand allocation is preferable to these or other contracting methods. However, we point out that these contracting methods require the buyer to possess significant bargaining power over the servers—the buyer must be able to control the pricing schedule used and its parameters, whereas demand allocation can be implemented by the buyer even if the buyer has little bargaining power. Therefore, because allocation policies are simple to implement and observed in practice, we suspect they are desirable vis-à-vis other techniques along at least some dimensions. Overall, additional research is needed to identify the situations in which demand allocation is the best option for the buyer.

In our analysis we assume the suppliers have identical cost functions, which is reasonable in markets that have homogeneous technologies. This naturally leads to symmetric equilibria. With heterogeneous cost functions, equilibrium analysis is more challenging. Zhang (2004) provides some initial results and finds that existence is less likely as costs become more heterogeneous: As one supplier gains a cost advantage it becomes necessary to dampen the competition among the suppliers to prevent one supplier from driving the other supplier out of the market, just like the problem we see when R is too high in the symmetric cost case. Hence, performance-based allocation of demand appears to be most effective when suppliers have comparable costs.

Our model assumes that each supplier only serves the buyer, as in the case when a supplier builds or reserves dedicated capacity for the buyer. In some cases each supplier may cater to multiple buyers, thereby creating two strategic decisions for each supplier: how much capacity to build and how to prioritize that capacity across buyers. Furthermore, there may be different prices for different priorities. The analysis of these systems is clearly beyond the scope of this research, but we again suspect that the buyer could use a smartly designed allocation policy to obtain higher priority from suppliers.

We conduct our analysis in the context of a queueing system, but there are also situations that may be better modeled as an inventory system: e.g., each supplier could choose a base-stock policy and the buyer is concerned with some dimension of the supplier's delivery lead-time distribution. While the specifics of the analysis would differ, we suspect that demand allocation would again be a useful tool for the buyer to motivate for better reliability among her suppliers.

Our analysis is conducted exclusively in steady state. For example, we assume that the buyer is able to infer each server's capacity from the servers' delivery times so that the correct demand share can be implemented. In practice the buyer would only obtain an estimate of each server's capacity. The significance of sampling error on our results is an open question.

Finally, we have implicitly assumed that the buyer is able to credibly commit to implement the chosen allocation policy. Without that ability, the buyer's set of allocation policies to choose from is quite limited. For example, if the buyer must implement a

state-independent policy, then only Bell-Stidham allocation is credible because it minimizes the buyer's waiting time for any set of capacities chosen. If the buyer implements a state-dependent policy, then only threshold allocation is credible, but not necessarily the same threshold policy discussed in §3.2. Again, the threshold policy must be chosen so as to minimize the buyer's waiting time for any capacity vector. Hence, the ability to credibly commit to an allocation policy is important to the buyer. We note that this same issue occurs in many other settings. For example, in the supply chain contracting literature, many coordinating contracts are studied and observed that require commitments: a buy-back contract is an a priori commitment by a supplier to pay a retailer for units returned by the retailer after stochastic demand occurs even though the supplier has no ex post incentive to do so.

5. Conclusion

In this paper, two queueing servers strategically choose their capacities/processing rates in response to a buyer's demand allocation policy. The buyer's objective is to design the allocation policy to achieve the shortest possible average delivery time from the servers, either by motivating the servers to build more capacity or by ensuring that the available capacity is effectively utilized. We focus on allocation policies based on the servers' capacities because we show that lead time-based allocation policies may not perform as well.

Previous research suggests that there may exist a trade-off between incentives and efficiency: An allocation policy that efficiently utilizes the servers' capacity may provide weak incentives for them to work quickly, and an allocation policy with strong incentives to work quickly may not effectively utilize the servers' capacity. We indeed demonstrate that there is considerable variation in the performance across allocation policies. Many allocation policies either provide absolutely no incentive for the servers to deliver quickly or provide too much competition among servers, thereby leading to unpredictable behavior. Even policies that are optimal for the buyer with nonstrategic and symmetric servers can perform poorly with strategic servers. However, we show that there need not be a trade-off between incentives and efficiency, i.e., there exists an allocation policy, threshold allocation, that induces the servers to work at their maximum rate and minimizes the buyer's lead time, given the resulting capacities.

Unfortunately, threshold allocation is complex. For example, its optimal parameters cannot be determined in closed form. We offer linear allocation as an alternative. Linear allocation also induces the servers to work at the maximum possible rate, but linear allocation does not utilize the servers' capacity as effectively as threshold allocation. In particular, because linear allocation allocates jobs immediately upon arrival and the assignment of jobs does not depend on the current state of the system (it is a state-independent allocation policy), linear allocation may allocate a job to a busy server while the other server is idle. Nevertheless, we show that linear and threshold allocations converge in performance at high utilizations, which suggests that linear allocation is attractive along many dimensions.

To conclude, a buyer should not ignore demand allocation as a strategy to obtain faster service, especially given its simplicity: There is no need to negotiate new contract terms or pricing with the servers because demand allocation can be implemented by a buyer without the explicit consent of the servers. However, creating competition among servers via their past performance requires some sophistication; a haphazard application of this strategy could have little impact.

Acknowledgments

The authors thank Saif Benjaafar, Noah Gans, Martin Lariviere, William Lovejoy, Serguei Netessine, Erica Plambeck, and Yong-Pin Zhou for their helpful comments, as well as the seminar participants at Columbia, Cornell, New York University, Northwestern University, University of California at Irvine, University of Minnesota, University of Washington, the Second MIT Symposium in Operations Research: Procurement and Pricing Strategies to Improve Supply Chain Performance, and the Competition with Delays meeting at Washington University. An electronic version of this paper is available from the authors' webpages. The previous version of this paper was titled "Procuring Fast Delivery, Part I: Multi-sourcing and Scorecard Allocation of Demand via Past Performance."

Appendix

Proof of Theorem 1. Server i's profit function is $\pi_i(\mu_i) = R\lambda_i - c(\mu_i)$. Let μ_i^0 be defined such that $\mu_i^0 > 0$ and $\lambda_i(\mu_i^0) = 0$ or $\mu_i^0 = 0$. $\pi_i(\mu_i)$ is then concave and decreasing for $\mu_i \in [0, \mu_i^0]$. Now differentiate π_i,

$$\frac{\partial \pi_i(\mu_i)}{\partial \mu_i} = R\left(1 - \frac{\mu_i^{1/2}}{\sum \mu_j^{1/2}} - \frac{\left(\sum \mu_j - \lambda\right)\mu_j^{1/2}}{2\mu_i^{1/2}\left(\sum \mu_j^{1/2}\right)^2}\right) - c'(\mu_i)$$

and, for notational convenience, let $B = \mu_j - \lambda$,

$$\frac{\partial^2 \pi_i(\mu_i)}{\partial \mu_i^2} = \frac{R\mu_j^{1/2}}{4\left(\sum \mu_j^{1/2}\right)^3}\left(\mu_j^{1/2} B \mu_i^{-3/2} + 3B\mu_i^{-1} - 3\mu_j^{1/2}\mu_i^{-1/2} - 1\right) - c''(\mu_i).$$

Define $f(\mu_i) = B\mu_j^{1/2}\mu_i^{-3/2} + 3B\mu_i^{-1} - 3\mu_j^{1/2}\mu_i^{-1/2} - 1$. If $B \le 0$, then $\pi_i''(\mu_i) \le 0$ and $\pi_i(\mu_i)$ is concave. Otherwise, it can be shown that $df/d\mu_i = 0$ has only one positive solution. Moreover, $f \to \infty$ and $df/d\mu_i < 0$ as $\mu_i \to 0$ and $f < 0$ as $\mu_i \to \infty$. Thus, f decreases from the positive domain to the negative

domain. Because $c'''(\mu_i) \geq 0$, there exists a $\mu_i^1 \geq \mu_i^0$ such that $\pi_i(\mu_i)$ is concave and decreasing for $\mu_i \in [0, \mu_i^0]$, convex for $\mu_i \in [\mu_i^0, \mu_i^1]$, and concave for $\mu_i^1 < \mu_i$. Because $\pi_i(0) = 0$, it follows that any interior solution to server i's first-order condition is a global optimum if at that solution profit is nonnegative.

The following equation provides the unique solution to the first-order conditions given the constraint $\mu_i = \mu_j$:

$$\left(\frac{R}{4}\right)\left(1 + \frac{\lambda/2}{\mu_i}\right) - c'(\mu_i) = 0.$$

(Because $c(\mu_i)$ is convex, the left-hand side is decreasing, so there is a unique solution.) The lower bound on R ensures that $\mu_{bs} > \lambda/2$. The condition $\pi_i(\mu_{bs}) \geq 0$ ensures that μ_{bs} is indeed a global optimum for all servers. □

Proof of Theorem 2. There are two significant complications to this analysis that prevent the use of standard existence and uniqueness results. (1) π_i is not unimodal (if $\mu_2 > \lambda$, then π_1 is concave and decreasing for $\mu_1 \in [0, \mu_2 - \lambda]$ and concave for $\mu_1 > \mu_2 - \lambda$, but not globally concave), which may create a discontinuity in the servers' best reply functions. (2) π_i is not differentiable at $\mu_i = \lambda - \mu_j$, which prevents the unconditional use of first-order conditions to determine the global maximizer of π_i. Let's first establish when $\{\mu_b, \mu_b\}$ is a Nash equilibrium under the given conditions. The servers' first-order conditions are satisfied when $c'(\mu_b) = R/2$, which yields a finite lead time only if $\mu_b > \lambda/2$, which simplifies to the first condition. However, because $\mu_i = 0$ can be optimal for a server, μ_b is an optimal response only if $\pi_i(\mu_b, \mu_b) \geq 0$, i.e., if $R\lambda/2 \geq c(\mu_b)$, which can be written as $c'(\mu_b) \geq c(\mu_b)/\lambda$ (the second condition). Now let's rule out other equilibria. Suppose $\{\mu_i, \mu_j\}$ is an equilibrium, $\mu_i \geq \mu_j$. Several cases need to be considered. (i) $\mu_i \geq \lambda + \mu_j$. If $\mu_j > 0$, then server j earns a negative profit, so this is not an equilibrium. If $\mu_j = 0$, then it must be that $\mu_i = \lambda$. For server j we have $\pi_j(\lambda, \mu_b) = c'(\mu_b)\mu_b - c(\mu_b) > 0$, breaking the equilibrium. (ii) $\mu_i + \mu_j < \lambda$, which implies $\mu_j < \lambda/2$. From server j's first-order condition we get $c'(\mu_j) = R > 2c'(\lambda/2)$, which implies $\mu_j > \lambda/2$ because $c'(\cdot)$ is increasing. Hence, we have a contradiction, so no equilibrium. (iii) $\mu_i + \mu_j = \lambda$. For this to be optimal for both servers it must be that $R \leq 2c'(\mu_i)$ and $R \leq 2c'(\mu_j)$, which cannot both be satisfied because $R > 2c'(\lambda/2)$. (iv) $\mu_i + \mu_j > \lambda$ and $\mu_i < \lambda + \mu_j$. Now the only solution to the first-order conditions is $\{\mu_b, \mu_b\}$.

To obtain an equilibrium with finite lead times, we need $\mu_i + \mu_j > \lambda$. Because $\mu_i \geq \lambda + \mu_j$ cannot be an equilibrium, there must be $\mu_i < \lambda + \mu_j$, which implies that $\{\mu_b, \mu_b\}$ is the only solution to the first-order conditions. However, $\{\mu_b, \mu_b\}$ is not an equilibrium if $c'(\mu_b) < c(\mu_b)/\lambda$, and $\mu_b < \lambda/2$ if $R < r_2$. □

Proof of Theorem 3. (i) For Nash equilibrium we need to show that μ_l maximizes server i's profit if the other server chooses $\mu_j = \mu_l$. The primary complication is due to the revenue term, $R\lambda_i(\mu)$, in the profit function. Server i's allocation is

$$\begin{aligned}\lambda_i(\mu) &= \min\{(\theta\mu_i/2 - \theta\mu_l/2 + \lambda/2)^+, \lambda\} \\ &= \min\left\{\left(\theta\mu_i/2 - \frac{\lambda}{2}\left(\frac{c'(\mu_l)\mu_l}{c(\mu_l)} - 1\right)\right)^+, \lambda\right\}.\end{aligned}$$

The second term is negative, so there exists some μ_i^0 such that $\lambda_i = 0$ for all $\mu_i \leq \mu_i^0$. If $\mu_i = \mu_l$, then $\lambda_i(\mu) = \lambda/2$. The condition $R > r_1$ ensures that $\lambda/2 < \mu_l$, so it follows that $\mu_i^0 < \mu_l$. The server's profit function is concave and decreasing for $\mu_i \leq \mu_i^0$ and concave and continuous for $\mu_i > \mu_i^0$, although possibly not differentiable when $\lambda_i(\mu) = \lambda$. For $\mu_i > \mu_i^0$, by construction of the parameters, π_i is maximized with $\mu_i = \mu_l$ and $\pi_i = 0$ with $\mu_i = \mu_l$. Therefore, μ_l is optimal for server i. Lead times are finite because $\mu_l > \lambda/2$. Next we concentrate on uniqueness.

Suppose μ is a Nash equilibrium of the capacity game. The proof first rules out asymmetric equilibrium with positive capacity for all servers and then equilibrium with $\mu_i = 0$ for some i are ruled out.

Suppose in some equilibrium $\mu_i > 0$ for all i. It must be, then, that $\lambda_i > 0$ for all i (otherwise server i would make negative profit). Thus, the first-order condition for each server must be satisfied given $\hat{n} = 2$, but that yields $\mu_i = \mu_l$ for all i because the solution to each server's first-order condition depends only on μ_i.

Now suppose there exists a $\mu_i = 0$ in equilibrium. All servers choosing $\mu_i = 0$ cannot be an equilibrium because then one server could build a small amount of capacity and earn positive profit. If server 1 has the only positive capacity, then server 1 receives $\lambda_1(\mu) = \lambda$ as long as $\mu_1 > 0$; this cannot be an equilibrium because then server 1's optimal capacity is some arbitrarily small capacity, which then allows the other server to build positive capacity and earn positive profit. (The condition $\mu_i > 0$ for all $i \leq \hat{n}$ in the allocation function is critical to this result.)

(ii) The proof is similar to (i), so it is omitted. □

Proof of Theorem 4. For server i

$$\frac{\partial \pi_i(\mu)}{\partial \mu_i} = R\lambda\beta\frac{\mu_i^{\beta-1}\mu_j^{\beta}}{\left(\sum_j \mu_j^{\beta}\right)^2} - (2a\mu_i + b), \quad i = 1, 2.$$

Given R and β, simple algebra reveals that $\mu_i = \mu_p$ is a symmetric solution to the first-order conditions, and it is the only solution. If each server chooses μ_p, then each server earns a zero profit, so it is an equilibrium if $\max \pi_i = 0$. Differentiate:

$$\frac{\partial^2 \pi_i(\mu)}{\partial \mu_i^2} = R\lambda\beta\frac{\mu_i^{\beta-2}\mu_j^{\beta}}{(\mu_i^{\beta} + \mu_j^{\beta})^3}[(\beta - 1)\mu_j^{\beta} - (\beta + 1)\mu_i^{\beta}] - 2a.$$

Note that $\beta = 2\bar{\mu}c'(\bar{\mu})/c(\bar{\mu}) > 2$. It can be shown (see Zhang 2004 for details) that π_i is concave-convex-concave if $\mu_j = \mu_p$. Because $\pi_i = 0$ and $\pi_i' < 0$ when $\mu_i = 0$, it must be that $\max \pi_i = 0$. Therefore, $\mu_1 = \mu_2 = \mu_p$ is the unique Nash equilibrium of the capacity game. □

Proof of Theorem 5. First demonstrate that $\mu_i = \mu_p$ for all i is a unique Nash equilibrium when $R > r_2$. The first-order conditions must be satisfied:

$$\frac{\partial \pi_i(\mu)}{\partial \mu_i} = R\lambda\frac{\mu_j}{\left(\sum_j \mu_j\right)^2} - c'(\mu_i) = 0, \quad i = 1, 2. \tag{9}$$

The first-order conditions imply $\mu_i = \mu_j$, so the only solution must be $\mu_i = \mu_p$ for all i. The condition $R > r_2$ ensures that the solution to the first-order conditions has $\mu_i > \lambda/2$, which provides finite lead times. Now suppose $R \leq r_2$ and there is an equilibrium with finite lead times. If the lead times are finite, then the first-order conditions (9) hold and only $\mu_i = \mu_j$ satisfy them. Again, because lead times are finite, it must

be that $\mu_i = \mu_j > \lambda/2$. Because each first-order condition is increasing in R, each first-order condition is maximized with $R = r_2$, in which case (9) can be written as

$$c'(\lambda/2)\lambda/(2\mu_i) - c'(\mu_i) < 0,$$

which means that $\mu_i = \mu_j$ cannot be an equilibrium. □

Proof of Lemma 6. See Kalai et al. (1992). □

Proof of Lemma 7. It is easy to determine that $z(R) = \mu_l/(\mu_l + \lambda/2)$, and the results follow immediately given that $\mu_l \geq \lambda/2$. □

Proof of Lemma 8. Each generates the maximum capacity, so $c(\mu_l) = R_l(\lambda/2)$ and $c(\mu_t) = R_t(\lambda/2)$, which imply $R_l/R_t = c(\mu_l)/c(\mu_t)$. If waiting times are equivalent, then

$$\frac{\mu_t}{\mu_t^2 - (\lambda/2)^2} = \frac{1}{\mu_l - \lambda/2},$$

which simplifies to $\mu_l/\mu_t = 1 - \rho^2 + \rho$. □

Proof of Lemma 9. Server i's profit is

$$\pi_i(W) = \lambda_i(W)R - c(\mu_i(W_i, \lambda_i(W))).$$

Let W be an equilibrium with $\mu_1(W_1, \lambda_1(W)) > \hat{\mu}$. Define $W_\varepsilon = (W_1 + \varepsilon, W_2)$ and $\lambda_\varepsilon = \lambda_1(W) - \lambda_1(W_\varepsilon)$: Server 1 is allocated λ_ε fewer units of demand because of the slower service provided to the buyer. Next we show that there exists an $\varepsilon > 0$ that would increase server 1's profit, which leads to a contradiction. The difference between the profit functions can be written as

$$\pi_1(W) - \pi_1(W_\varepsilon) = \lambda_\varepsilon R - [c(\mu_1(W_1, \lambda_1(W))) - c(\mu_1(W_1 + \varepsilon, \lambda_1(W_\varepsilon)))].$$

We know that $\mu_1(W_1, \lambda_1(W)) > \hat{\mu}$ implies $c'(\mu_1(W_1, \lambda_1(W))) > R$. Because λ_1 is continuous for a sufficiently small ε, there is also $c'(\mu_1(W_1 + \varepsilon, \lambda_1(W_\varepsilon))) > R$. By assumption,

$$\lambda_\varepsilon < \mu_1(W_1, \lambda_1(W)) - \mu_1(W_1 + \varepsilon, \lambda_1(W_\varepsilon)),$$

i.e., the required capacity decreases more than demand when the server provides worse service, we know that

$$\lambda_\varepsilon R < c(\mu_1(W_1, \lambda_1(W))) - c(\mu_1(W_1 + \varepsilon, \lambda_1(W_\varepsilon)))$$

or $\pi_1(W) - \pi_1(W_\varepsilon) < 0$. Therefore, there exists an ε that would increase server 1's profit. As a result, W cannot be an equilibrium. □

References

Allon, G., A. Federgruen. 2003. Competition in service industries. *Oper. Res.* Forthcoming.

Anton, J. J., D. A. Yao. 1989. Split awards, procurement, and innovation. *RAND J. Econom.* **20**(4) 528–552.

Anton, J. J., D. A. Yao. 1992. Coordination in split award auctions. *Quart. J. Econom.* **107**(2) 681–707.

Anupindi, R., R. Akella. 1993. Diversification under supply uncertainty. *Management Sci.* **39**(8) 944–963.

Armony, M., E. Plambeck. 2005. The impact of duplicate orders on demand estimation and capacity investment. *Management Sci.* **51**(10) 1505–1518.

Bell, C., S. Stidham. 1983. Individual versus social optimization in the allocation of customers to alternative servers. *Management Sci.* **29**(7) 831–839.

Benjaafar, S., E. Elahi, K. L. Donohue. 2007. Outsourcing via service competition. *Management Sci.* **53**(2) 241–259.

Bernstein, F., G. DeCroix. 2004. Decentralized pricing and capacity decisions in a multitier system with modular assembly. *Management Sci.* **50**(9) 1293–1308.

Bernstein, F., F. de Vericourt. 2005. Competition for procurement contracts with service guarantees. *Oper. Res.* Forthcoming.

Bonser, J. S., S. D. Wu. 2001. Procurement planning to maintain both short-term and adaptiveness and long-term perspective. *Management Sci.* **47**(6) 769–786.

Cachon, G. P., P. T. Harker. 2002. Competition and outsourcing with scale economies. *Management Sci.* **48**(10) 1314–1333.

Cachon, G. P., M. A. Lariviere. 1999a. Capacity allocation using past sales: When to turn and earn. *Management Sci.* **45**(5) 685–703.

Cachon, G. P., M. A. Lariviere. 1999b. Capacity choice and allocation: Strategic behavior and supply chain performance. *Management Sci.* **45**(8) 1091–1108.

Cachon, G. P., M. A. Lariviere. 1999c. An equilibrium analysis of linear, proportional and uniform allocation of scarce capacity. *IIE Trans.* **31**(9) 835–850.

Cachon, G. P., F. Zhang. 2006. Procuring fast delivery: Sole-sourcing with information asymmetry. *Management Sci.* **52**(6) 881–96.

Cachon, G. P., P. H. Zipkin. 1999. Competitive and cooperative inventory policies in a two-stage supply chain. *Management Sci.* **45**(7) 936–953.

Chayet, S., W. Hopp. 2002. Lead time competition under uncertainty. Working paper, Northwestern University, Evanston, IL.

Chen, J., D. D. Yao, S. Zheng. 2001. Optimal replenishment and rework with multiple unreliable supply sources. *Oper. Res.* **49**(3) 430–443.

Christ, D., B. Avi-Itzhak. 2002. Strategic equilibrium for a pair of competing servers with convex cost and balking. *Management Sci.* **48**(6) 813–820.

Deshpande, V., L. Schwarz. 2002. Optimal capacity choice and allocation in decentralized supply chains. Working paper, Purdue University, West Lafayette, IN.

Elmaghraby, W. J. 2000. Supply contract competition and sourcing strategies. *Manufacturing Service Oper. Management* **2**(4) 350–371.

Farlow, D., G. Schmidt, A. Tsay. 1996. Supplier management at Sun Microsystems (A). Stanford Business School Case, Stanford University, Stanford, CA.

Gallien, J., L. M. Wein. 2005. A smart market for industrial procurement with capacity constraints. *Management Sci.* **51**(1) 76–91.

Gilbert, S. M., Z. K. Weng. 1998. Incentive effects favor nonconsolidating queues in a service system: The principle-agent perspective. *Management Sci.* **44**(12) 1662–1669.

Ha, A. Y., L. Li, S.-M. Ng. 2003. Price and delivery logistics competition in a supply chain. *Management Sci.* **49**(9) 1139–1153.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA.

Kalai, E., M. I. Kamien, M. Rubinovitch. 1992. Optimal service speeds in a competitive environment. *Management Sci.* **38**(8) 1154–1163.

Koole, G. 1995. A simple proof of the optimality of a threshold policy in a two-server queuing system. *Systems Control Lett.* **26** 301–303.

Lederer, P. J., L. Li. 1997. Pricing, production, scheduling, and delivery time competition. *Oper. Res.* **45**(3) 407–420.

Li, C.-L., P. Kouvelis. 1999. Flexible and risk-sharing supply contracts under price uncertainty. *Management Sci.* **45**(10) 1378–1398.

Li, L. 1992. The role of inventory in delivery-time competition. *Management Sci.* **38**(2) 182–197.

Lin, W. P., R. Kumar. 1984. Optimal control of a queuing system with two heterogeneous servers. *IEEE Trans. Automatic Control* **29** 696–703.

Martinez de Albeniz, V., D. Simchi-Levi. 2003. Competition in the supply option market. Working paper, MIT, Boston, MA.

Minner, S. 2003. Multiple-server inventory models in supply chain management: A review. *Internat. J. Production Econom.* **81–82** 265–279.

Nagarajan, M., Y. Bassok. 2003. A bargaining framework in supply chains. Working paper, University of Southern California, Los Angeles, CA.

Naor, P. 1969. On the regulation of queue size by levying tolls. *Econometrica* **37** 15–24.

Pyke, D., E. Johnson. 2003. Sourcing strategy and supplier relationships: Alliances vs. Eprocurement. C. Billington, H. Lee, J. Neale, T. Harrison, eds. *The Practice of Supply Chain Management*. Kluwer Publishing, Boston, MA, 77–89.

Rubinovitch, M. 1985a. The slow server problem. *J. Appl. Probab.* **22** 205–213.

Rubinovitch, M. 1985b. The slow server problem: A queue with stalling. *J. Appl. Probab.* **22** 879–892.

Sedarage, D., O. Fujiwara, H. T. Luong. 1999. Determining optimal order splitting and reorder level for N-server inventory systems. *Eur. J. Oper. Res.* **116** 389–404.

Seshadri, S. 1995. Bidding for contests. *Management Sci.* **41**(4) 561–576.

Seshadri, S., K. Chatterjee, G. L. Lilien. 1991. Multiple source procurement competitions. *Marketing Sci.* **10**(3) 246–263.

So, K. C. 2000. Price and time competition for service delivery. *Manufacturing Service Oper. Management* **2**(4) 392–409.

Talluri, S. 2002. A buyer-seller game model for selection and negotiation of purchasing bids. *Eur. J. Oper. Res.* **143**(1) 171–180.

Walgrand, J. 1984. A note on "Optimal control of a queuing system with two heterogeneous servers." *Systems Control Lett.* **4** 131–134.

Wang, Y., Y. Gerchak. 2003. Capacity games in assembly systems with uncertain demand. *Manufacturing Service Oper. Management* **5**(3) 252–267.

Yoffie, D. B. 1990. The global semiconductor industry, 1987. D. B. Yoffie, ed. *International Trade and Competition: Cases and Notes in Strategy and Management*. McGraw-Hill, New York, 389–413.

Zhang, F. 2004. Coordination of lead times in supply chains. Dissertation, University of Pennsylvania, Philadelphia, PA.