M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 12, No. 3, Summer 2010, pp. 511–526
ISSN 1523-4614 | EISSN 1526-5498 | 10 | 1203 | 0511

informs

DOI 10.1287/msom.1090.0282

Exploiting Market Size in Service Systems

Sunil Kumar
Graduate School of Business, Stanford University, Stanford, California 94305, skumar@stanford.edu

Ramandeep S. Randhawa
Marshall School of Business, University of Southern California, Los Angeles, California 90089,
ramandeep.randhawa@marshall.usc.edu

We study a profit-maximizing firm providing a service to price and delay sensitive customers. We are interested in analyzing the scale economies inherent in such a system. In particular, we study how the firm's pricing and capacity decisions change as the scale, measured by the potential market for the service, increases. These decisions turn out to depend intricately on the form of the delay costs seen by the customers; we characterize these decisions up to the dominant order in the scale for both convex and concave delay costs. We show that when serving customers on a first-come, first-served basis, if the customers' delay costs are strictly convex, the firm can increase its utilization and extract profits beyond what it can do when customers' delay costs are linear. However, with concave delay costs, the firm is forced to decrease its utilization and makes less profit than in the linear case. While studying concave delay costs, we demonstrate that these decisions depend on the scheduling policy employed as well. We show that employing the last-come, first-served rule in the concave case results in utilization and profit similar to the linear case, regardless of the actual form of the delay costs.

Key words: service systems; pricing; capacity planning; large market size; nonlinear delay costs; convex delay costs
History: Received: September 22, 2008; accepted: September 14, 2009. Published online in *Articles in Advance* December 29, 2009.

1. Introduction

Increasing the scale of a service operation, by serving a larger market perhaps, has several benefits for the service provider. Given that service operations typically face variability, and serve customers who care about the impact of this variability (as measured by delay), the provider can mitigate this impact by exploiting statistical economies of scale inherent in such operations. This exploitation can take the form of increasing the utilization of the service, by making the realized traffic at the service heavier, without degrading quality as perceived by its customers. It can also take the form of increasing prices to extract the surplus provided to customers in the form of better quality of service. And it can be any combination of these two approaches. The primary goal of this paper is to estimate the limits of the benefits of increased scale to a monopolistic service provider. In particular, we want to know how the system utilization increases and how much profit the firm can expect to make as the scale increases. Moreover, we want to know how these limits depend on the customers' price and delay sensitivity, as well as the configuration by which the service is provided.

Perhaps the simplest example by which to observe such benefits of scale is in staffing a call center, where pricing is not a lever that can be exploited. If the customers experience linear disutility to delay, or in other words, linear delay costs, and the provider experiences linear staffing costs, the optimal prescription is the well-known square root staffing rule. When facing a high volume of customer calls, the provider staffs so that the capacity is equal to the arrival rate plus a "safety capacity" proportional to the square-root of the arrival rate to handle variability (see the survey paper by Gans et al. 2003 for a detailed description). Under this rule, if the scale (as measured by the arrival rate) increases fourfold, the safety capacity only doubles; that is, as the scale increases, the safety capacity expressed as a fraction of the scale decreases, eventually approaching zero. Despite this, increasing the scale of the system ensures that the quality

of service actually improves; the delays observed by the customers become smaller. This combination of increasingly high utilization and increasingly high quality of service illustrate the benefit of increasing scale to the provider.

The square-root staffing rule is known to be the best the provider can do as long as customers value delay linearly. In general, this need not be the case. Customers could value delay either in a convex fashion, disliking the second unit of delay more than the first, or in a concave fashion, disliking the first unit of delay more than a subsequent unit of delay. There is no reason to pick one or the other model of customers' delay sensitivity a priori. Moreover, pricing the service is also an important issue. The goal of this paper is to study how a provider can exploit increasing scale in a market of price and delay sensitive customers, under both convex and concave delay sensitivity. In our setting, increasing scale translates to increasing the potential market of homogenous customers. If price and the distribution of delay were both held fixed, then increasing the scale simply translates to an identical increase in arrival rate; that is, in our setting, each customer's behavior is independent of the scale of the operation the customer is served by. This is a natural assumption for the question we are interested in, namely, how does growing the customer base benefit the provider?

Arguing by analogy from the square-root staffing rule, we can conjecture that as scale increases, the safety capacity as a fraction of the scale decreases and the delay improves. Adding pricing complicates the picture, but the basic intuition does go through. However, the nonlinearity of delay sensitivity can also be exploited. Consider the convex sensitivity case. Because customers with convex sensitivity are less sensitive to small delays than customers with linear sensitivity, and we expect delays to be small when the scale is large, the provider may actually do better in the convex case than the linear case. Our analysis bears out this intuition and precisely characterizes the best the provider can do. A by-product of our analysis shows that the provider operates in a regime where the configuration by which the service is provided (one fast server or many slow servers) is irrelevant up to the dominant order. We also obtain the analogous but opposite result for the concave sensitivity case; the provider does worse than in the linear case if she chooses to serve customers on a first-come, first-served (FCFS) basis. Although the FCFS service discipline is the choice for linear or convex delay sensitivity, it is not the correct choice for concave sensitivity, where the opposite, last-come, first-served (LCFS), is the preferred discipline. When LCFS discipline is used, we find that the best the provider can do is the same as in the linear case; somewhat surprisingly, the service rule negates the effect of concave delay sensitivity.

Finally, a note on the motivation for asymptotic analysis in our paper is in order. In many papers that study queueing in service operations, asymptotic analysis is used as an analytical tool; that is, the papers are really interested in providing the answer to one particular problem, and they find it convenient to embed that problem in a sequence of problems whose asymptotic limits are easier to characterize. This allows for considerable leeway in how the sequence is chosen. In contrast, in our setting, we are actually interested in the behavior of one particular sequence, which is predetermined; that is, our sequence is always the one corresponding to increasing the scale as measured by market size, holding everything else fixed. Moreover, the sequence is not just an analytical device to arrive at an answer conveniently; it is the answer to our primary question, namely, how can the provider exploit scale?

1.1. Literature Review

There is a vast amount of literature devoted to the use of different forms of the square-root rules in large systems. The basis for this literature is the heavy traffic approximation of queueing systems, which can be broadly divided into two streams: one that deals with a single server or a fixed number of servers (see, for example, Kingman 1961), and one that deals with the so-called "many-server" systems, introduced in Halfin and Whitt (1981). These two modeling approaches have a fundamental difference. The analysis of single-server queues only requires that the utilization be near 100% and uses a time scaling approach to derive the approximations. However, the many-server regime requires the system size, or the number of servers, to increase without bound as well

as the utilization to approach 100%; the system size provides the natural operating scale of the system.

The approximations derived by both these methods have found great applicability in computing performance estimates in queueing systems, and computing excellent routing or scheduling policies. In single-server systems, for example, Van Mieghem (1995), Stolyar (2004), and Mandelbaum and Stolyar (2004) solve scheduling problems with convex delay costs; Ata and Olsen (2009) solve similar problems with convex-concave costs. The many-server regime has primarily been used to analyze staffing and scheduling in call centers; for example, see Garnett et al. (2002), Armony (2005), Dai and Lin (2008), Gurvich and Whitt (2008), and Tezcan (2008). These papers consider a square-root staffing regime where the number of servers is of the form $\alpha n + \beta\sqrt{n}$, where n denotes the system size (usually market size is measured by the customer arrival rate).

Recently, there has been a shift toward utilizing the aforementioned heavy traffic approximation machinery to studying queueing systems from an economic perspective, i.e., viewing the customers as rational and the service provider as profit maximizing.[1] In such settings, one needs to verify that the profit is indeed maximized when operating in the square-root regime. Some examples of papers that provide such justification are Maglaras and Zeevi (2003), Maglaras and Zeevi (2005), Plambeck and Ward (2006), and Randhawa and Kumar (2008). These papers use scale-independent delay costs similar to those in this paper; however, they restrict attention to costs that are linear in the performance measure. On the other hand, there are papers that deal with nonlinear delay costs, without including pricing, that perform their analysis by scaling these costs; see, for example, Van Mieghem (1995), Stolyar (2004), Mandelbaum and Stolyar (2004), Dai and Lin (2008), and Gurvich and Whitt (2008). When dealing with nonlinear costs, one can choose to incorporate the effect of the entire delay cost function by "scaling" the delay cost. However, such a "cost scaling" approach would require that customers *alter* their behavior according to the size of the system. This may not be appropriate from a modeling perspective, especially in systems with rational customers. Therefore, our approach is to simply *not* scale costs, that is, we assume that the rational customers' delay sensitivity is absolute and independent of the system size. There is a dearth of literature that undertakes such "scale-independent" analysis. In fact, Whitt (2003) and Borst et al. (2004) are the only papers we are aware of that follow this approach. Filling this lacuna while incorporating pricing is the primary focus of our paper.

Though this paper is related to Whitt (2003) and Borst et al. (2004), there are important differences. Whitt (2003) studies the impact of congestion on the performance of the system by positing an aggregate relationship between the delay in the system and the customer arrival rate. The ensuing equilibrium provides an illustration of a setting where the operating regime differs from the conventional square-root regime. In contrast, we explicitly model the relation between customer disutility and system delay and study the behavior of the firm's profit maximizing regime as the system size increases. Borst et al. (2004) derives asymptotically optimal capacity prescriptions with the objective of minimizing the sum of delay and capacity costs for many-server configurations. In contrast, the objective of this paper is to study the manner in which the optimal operating regime depends on the structure of scale-independent delay costs, and in particular how the firm can exploit the increasing scale by both pricing and capacity sizing. We find that for strictly convex delay costs that have a zero derivative at the origin, the optimal operating regime is heavier than conventional square-root regimes, and is one where the many-server configuration reduces to a single-server configuration. Our results further demonstrate that the optimal operating regime depends on the scheduling policy employed in addition to the delay costs. The choice of the scheduling policy is moot in the case of linear delay costs, and thus demonstrates that additional care must be taken when dealing with nonlinear delay costs.

1.2. Organization

We begin in §2 with a formal model description and a summary of results. Section 3 is the core of this paper and studies the joint pricing and capacity sizing problem, and derives the optimal strategy for a firm in

[1] Economic analysis of queueing systems by itself has a long history going back to Naor (1969). See Hassin and Haviv (2003) for a detailed review.

wake of increasing scale. Sections 4 and 5 consider the case where one decision variable is held constant and the other is optimized. In §4, we study a situation where the capacity growth is predetermined and only price has to be chosen. This case turns out to be more complicated than the general case, but it serves as a useful detour to illustrate when heavy traffic may not be achieved. Next, in §5, we consider the problem of optimizing capacity when prices are exogenously fixed. Sections 3–5 focus on the single-server configuration. All of the results hold in an appropriate sense in the many-server configuration as well, and this is the subject of §6. We conclude in §7. Proofs of all results are relegated to the appendix.

2. Model and Summary of Results

We consider a monopolistic firm that serves a market of price and delay sensitive customers. We model the firm as a queueing system, with the firm providing service either via a single-server or a many-server configuration, in an FCFS fashion. The service requirement of all customers are independent and exponentially distributed with the same mean. The firm sets an access price p and chooses either a service rate μ in the single-server setting, or the number of servers s in the many-server setting, to maximize its profit. The firm faces a linear cost of capacity.

Customers arrive to the firm as a Poisson process with rate $n\Lambda(p+\xi)$, where n represents the market size, $\Lambda(\cdot)$ represents a demand curve, and ξ represents the delay cost. For the most part, the crucial assumption in this paper is that the delay cost is of the form $\xi = h\mathbb{E}W^r$ for some $r \geq 1$ and $h > 0$, where W denotes the steady-state delay (time in queue). This demand function can be interpreted as follows: potential customers arrive to the firm as a Poisson process with rate $n\Lambda(0)$. Each potential customer has a valuation of the service V, which is independent and identically distributed across the customers. The potential customers have no system information, and thus decide to join the system only if their valuation exceeds the expected cost of joining, which is the sum of the price p and the expected steady-state delay cost ξ; that is, potential customers join the system only if $V \geq p+\xi$. Thus, the effective rate at which customers arrive to the firm is $n\Lambda(0)\mathbb{P}(V \geq p+\xi)$. Note that this demand function represents customers who are agnostic tothe size of the market—at a given p and ξ, their behavior is the same regardless of n; n simply scales the overall rate seen by the firm. This seems to be a reasonable assumption.

The system manager sets p and the capacity level to maximize the firm's expected profit. The profit function will be made explicit in each of the settings studied later.

In all the ensuing analyses, we make the following assumptions on the demand function.

Assumption 1. (a) $\Lambda\colon (0,\infty) \to (0,\infty)$ *is positive, decreasing, continuously differentiable, and satisfies*

$$\lim_{p\to 0}\Lambda(p) = \infty \quad \textit{and} \quad \lim_{p\to\infty}\Lambda(p) = 0.$$

(b) *The revenue function* $p\Lambda(p)$ *has no trivial maximizers, that is,* $\arg\max p\Lambda(p) \in (0,\infty)$.

The assumption that the domain and range of Λ is $(0,\infty)$ is for convenience alone and can easily be relaxed to the case when both the domain and range are intervals of the form (x,y) with $0 \leq x < y \leq \infty$ (which includes the case of linear demand).

Before continuing, we introduce one piece of notation: for real-valued functions $a, b\colon \mathbb{R} \to \mathbb{R}_+$, we use the notation $a(n) = \Theta(b(n))$ to denote the existence of constants $C_1, C_2 > 0$ such that $C_1 b(n) \leq a(n) \leq C_2 b(n)$ for all $n \in \mathbb{N}$. We also use the notation $a(n) = o(b(n))$ to denote $a(n)/b(n) \to 0$ as $n \to \infty$.

In this paper, we consider two configurations by which the firm provides service, namely, the single-server and the many-server configurations. For the most part, we focus on the single-server configuration. We will briefly discuss the case of the many-server configuration in §6.

2.1. Single-Server Configuration

In this configuration, customers are served by a single server; this is the well-known $M/M/1$ setting. The system manager's objective is to maximize the firm's profit by appropriately setting the price and capacity level measured by the service rate μ. The demand seen by the firm (i.e., the arrival rate as a function of the price and delay) is $n\Lambda(p+\xi)$. There is a cost of $\kappa\mu$ with $\kappa > 0$ associated with selecting a service rate μ. Thus, the firm's profit upon setting a price p and capacity level μ is $pn\Lambda(p+\xi) - \kappa\mu$. The system

manager's optimization problem in this setting can be written as

$$\begin{aligned} \max_{p,\,\mu\geq 0} \quad & pn\Lambda(p+\xi)-\kappa\mu \\ \text{s.t.} \quad & \mu > n\Lambda(p+\xi) \\ & \xi = h\mathbb{E}W^r. \end{aligned}$$

This optimization problem optimizes the firm's expected steady-state profit in *equilibrium*. In this case, for a given price and capacity pair, an equilibrium is defined by a steady-state delay distribution such that the corresponding expected delay cost ξ induces a time-homogenous arrival rate $n\Lambda(p+\xi)$ consistently. As the equilibrium arrival rate is time homogenous, we can use the well-known results for an $M/M/1$ queue to characterize ξ. In particular, for an $M/M/1$ queue with arrival rate λ, service rate μ, and utilization $\rho=\lambda/\mu<1$, the distribution of the steady-state delay is given by

$$\mathbb{P}(W>t)=\rho e^{-(\mu-\lambda)t}, \tag{1}$$

and hence the delay cost $h\mathbb{E}W^r$ is given by

$$h\mathbb{E}W^r = h\rho\frac{\Gamma(r+1)}{(\mu-\lambda)^r}, \tag{2}$$

where Γ denotes the Gamma function and satisfies $\Gamma(r+1)=r!$ for $r\in\mathbb{Z}_+$. We direct the reader to Asmussen (2003, §9 of Chapter III) for details. Thus, replacing λ by the equilibrium arrival rate $n\Lambda(p+\xi)$, we obtain

$$\xi\equiv h\mathbb{E}W^r = h\frac{n\Lambda(p+h\mathbb{E}W^r)}{\mu}\frac{\Gamma(r+1)}{(\mu-n\Lambda(p+h\mathbb{E}W^r))^r}. \tag{3}$$

The system manager's optimization problem can thus be written as

$$\begin{aligned} \max_{p,\,\mu\geq 0} \quad & pn\Lambda(p+\xi)-\kappa\mu \\ \text{s.t.} \quad & \mu > n\Lambda(p+\xi) \\ & \xi = h\frac{n\Lambda(p+\xi)}{\mu}\frac{\Gamma(r+1)}{(\mu-n\Lambda(p+\xi))^r}. \end{aligned} \tag{4}$$

2.2. Notion of k-Heavy Traffic and Summary of Results

Our first result is to establish that an increase in scale, under suitable conditions (that will be made precise), leads to a system utilization that is near 100%. We differentiate between the ways in which the utilization at market size n, denoted by ρ_n, approaches 1, i.e., between the different types of the so-called heavy traffic regimes. For this we introduce the notion of the k-heavy traffic regime.

Definition 1 (k-Heavy Traffic Regime). A system with traffic intensity ρ_n, where n is the market size, is said to be in a k-heavy traffic regime for $k\in\mathbb{R}_+$ if $n^k(1-\rho_n)\to C\in(0,\infty)$ as $n\to\infty$.

According to this definition, the conventional heavy traffic regime where $\sqrt{n}(1-\rho_n)$ converges to some positive, finite constant will be henceforth referred to as 1/2-heavy traffic.

In the view of achievable limits of the exploitation of market size, our main result relates the heavy traffic regime to the delay sensitivity of the customers. In particular, if customers value delay (or suffer disutility) as $\mathbb{E}W^r$, for $r\geq 1$, where W is the realized delay, then r determines the heavy traffic regime. The regime achieved by a rational firm is $r/(r+1)$-heavy traffic as defined above; that is, the larger the value of r, the heavier the traffic. Moreover, the maximal achievable profit is $\Pi_n^\star = n(\overline{\Pi}-\Theta(n^{-r/(r+1)}))$, where $\overline{\Pi}$ is an "unattainable ideal" profit (per unit of market size), which corresponds to no delays at all; that is, the firm achieves a profit that deviates from an unattainable ideal by a quantity that diminishes with r.

Turning now to achieving the limits of attainable performance identified above, we carry out an asymptotic analysis in the regime identified. We provide explicit prescriptions for price and capacity that achieve these limits up to a negligible tolerance. This provides a complete normative perspective on asymptotic analysis of such systems. As a by-product of our analysis, we show that for strictly convex costs ($r>1$) our prescriptions are the same, regardless of configuration (single or many server). This equivalence of configuration has not been explicitly noted before in literature.

The case when $r<1$, i.e., there are concave costs, is more involved. If the firm chooses to operate under the FCFS discipline, then results analogous to the $r\geq 1$ case are obtained. The maximal profit is $\Pi_n = n(\overline{\Pi}-\Theta(n^{-r/(r+1)}))$, and the firm operates in $r/(r+1)$-heavy traffic. However, it may not be in the firm's interest to operate FCFS. If, for example, it chooses to operate

LCFS, which is optimal for the single-server configuration, then the results are qualitatively different. The firm always operates in 1/2-heavy traffic, regardless of r, and the profit is $n\bar{\Pi} - \Theta(\sqrt{n})$. This allows us to conclude, somewhat surprisingly, that the choice of the scheduling rule may determine the type of heavy traffic achieved. This suggests caution especially with approaches that first fix the heavy traffic regime and then attempt to determine a scheduling rule.

3. Joint Pricing and Capacity Sizing: Limits of Achievable Performance and Prescriptions

This section focuses on solving the joint pricing and capacity sizing problem for different delay cost structures. For the most part, we focus on delay costs of the form $\mathbb{E}W^r$. Section 3.1 discusses the case of convex costs ($r \geq 1$), and §3.2 discusses the case of concave costs ($0 < r < 1$). Going beyond power functions, §3.3 discusses the case of general delay costs.

3.1. Convex Delay Costs ($r \geq 1$)

In this case, the system manager optimizes profits jointly on price and capacity by solving the optimization problem (4). Before analyzing the solution analytically, we discuss a numerical illustration of the interaction among the optimal solution, system scale, and delay cost structure. We consider the case of linear demand, $\Lambda(x) = 4 - x$ for $0 \leq x \leq 4$, and compare the optimal solution for linear and quadratic delay costs, i.e., $\xi = h\mathbb{E}W^r$ for $r = 1, 2$. We fix the parameters $h = 1$ and $\kappa = 1$.

Table 1 compares the optimal solution (which is computed as a numerical solution of (4)) for the case of linear and quadratic delay costs for different system sizes. As one expects for small system sizes ($n = 1$), the system with linear delay costs dominates the one with quadratic delay costs. However, as the system scale increases, quadratic delay costs lead to higher profits compared to the linear case. Note that the utilization is also higher for the system with quadratic delay costs. This is intuitive because when the system scale increases, the delay reduces due to the economies of scale in queueing systems. When the delay is very small, quadratic delay costs are much lower than linear delay costs, and this allows the system manager to increase utilization, and hence generate higher profits in the quadratic case.

Table 1 Comparison of the Optimal Solution for Linear and Quadratic Delay Costs

	Linear delay costs		Quadratic delay costs	
n	Traffic intensity	Profit	Traffic intensity	Profit
1	0.54	0.52	0.45	0.20
10	0.79	15.72	0.79	17.09
100	0.92	201.6	0.95	212.6

Notes. For a smaller scale ($n = 1$), the system with quadratic delay costs has a lower traffic intensity and profit. As the scale increases, the system with quadratic costs has a higher traffic intensity and profit.

We now formalize this intuition by analyzing (4). For convenience, we will use the standard method of rewriting (4) as a problem of selecting the optimal arrival rate instead of the optimal price. Note that given an arrival rate λ, the price that will "generate" this arrival rate in equilibrium solves

$$n\Lambda(p + \xi) = \lambda,$$

and hence, using the expression for delay cost in (2), it is given by

$$p = \Lambda^{-1}(\lambda/n) - h\frac{\lambda}{\mu}\frac{\Gamma(r+1)}{(\mu - \lambda)^r}. \tag{5}$$

Thus, the optimization problem (4) can be equivalently solved with the arrival rate as a decision variable instead of the price. The corresponding objective function is now given by $(\Lambda^{-1}(\lambda/n) - h(\lambda/\mu)(\Gamma(r+1)/(\mu - \lambda)^r))\lambda - \kappa\mu$. Next, we replace the dummy variable for arrival rate λ by $n\hat{\lambda}$, and that for capacity μ by $n\hat{\mu}$. (We will use "$\hat{\cdot}$" to represent scaled parameters throughout this paper.) This gives us the following equivalent problem:

$$n\left[\max_{\hat{\lambda} \geq 0,\, \hat{\mu} > \hat{\lambda}} \left(\Lambda^{-1}(\hat{\lambda}) - h\frac{\hat{\lambda}}{\hat{\mu}}n^{-r}\frac{\Gamma(r+1)}{(\hat{\mu} - \hat{\lambda})^r}\right)\hat{\lambda} - \kappa\hat{\mu}\right]. \tag{6}$$

Let $\Pi_n^\star$ and $(\hat{\lambda}_n^\star, \hat{\mu}_n^\star)$ denote the optimal objective function and any optimizer of this problem, respectively. We will lay out the large market asymptotic properties of the optimal profit $\Pi_n^\star$ by first deriving an upper bound on $\Pi_n^\star$ for all n. We will compute an upper bound that is natural, namely, the profit in an idealized system that incurs no delays at all, denoted by $\Pi_\infty^\star$. Then, we will estimate the deviation of $\Pi_n^\star$ from $\Pi_\infty^\star$.

It is straightforward to see that the profit in the idealized system $\Pi^\star_\infty$ is the solution to the following optimization problem:

$$\max_{\bar\lambda \geq 0} \Lambda^{-1}(\bar\lambda)\bar\lambda - \kappa\bar\lambda. \tag{7}$$

We assume that any optimizer of this ideal problem $\hat\lambda^\star_\infty$ satisfies $0 < \hat\lambda^\star_\infty < \infty$. Then, using the first-order conditions that the optimal solutions must satisfy, we obtain the following characterization.

Proposition 1. *For $r \geq 1$, under any solution of* (6)*, the system operates in $r/(r+1)$-heavy traffic. Furthermore, the optimal profit is given by $\Pi^\star_n = n[\Pi^\star_\infty - \Theta(n^{-r/(r+1)})]$.*

The proofs of all results are postponed to the appendix. This result states that the optimal operating regime is $r/(r+1)$-heavy traffic. Note that for $r > 1$, this implies that the optimal utilization is greater than 1/2-heavy traffic, which is optimal for the case $r = 1$. This is intuitive because for large system sizes, the delay will be negligible, and hence strictly convex costs ($r > 1$) lead to further lower delay costs compared to the linear case. This allows the system manager to increase utilization, and generate higher profits compared to the linear case.

Remark 1. The fact that the optimal solution leads to $r/(r+1)$-heavy traffic follows from the first-order optimality conditions, and an intuitive explanation is as follows. It is clear that the optimal capacity must satisfy $n\hat\mu^\star_n = n\hat\lambda^\star_n + \delta^\star_n$ for some (relatively) small $\delta^\star_n > 0$; that is, the rate at which delay cost is accumulated under the optimal solution is proportional to $n(1/(n\hat\mu^\star_n - n\hat\lambda^\star_n)^r) = n(\delta^\star_n)^{-r}$ (this follows from (2)). The optimal solution will, in fact, be such that neither the delay cost nor the cost of capacity in excess of the arrival rate dominate in the limit; that is, we must have $\Theta(n(\delta^\star_n)^{-r}) = \Theta(\delta^\star_n)$, and we obtain that $\delta^\star_n = \Theta(n^{1/(r+1)})$, i.e., the optimal operating regime is $r/(r+1)$-heavy traffic.

We now characterize the optimal prescriptions. We lower our aspirations to computing prescriptions that are within a negligible tolerance of the optimal solution. One approach is to use the first-order optimality conditions to characterize the optimal solution. Instead, we present a more generic approach that can potentially be extended to more complex systems.

We begin by noting that $(\hat\lambda^\star_n, \hat\mu^\star_n) \to (\hat\lambda^\star_\infty, \hat\lambda^\star_\infty)$ as $n \to \infty$, where $\hat\lambda^\star_\infty$ is the optimizer of (7); for convenience, we will assume that (7) has a unique solution. Thus, for our prescription, it suffices to consider solutions of the form $\hat\lambda_n = \hat\lambda^\star_\infty + \epsilon_n$ and $\hat\mu_n = \hat\lambda^\star_\infty + \delta_n$ with $\epsilon_n < \delta_n$, and $\epsilon_n, \delta_n \to 0$ as $n \to \infty$.

Using a Taylor series expansion around $\hat\lambda^\star_\infty$, we can write the corresponding profit function

$$\Pi_n(\hat\lambda_n, \hat\mu_n) = n\Bigg[\Pi^\star_\infty + \epsilon_n\Bigg[\frac{\hat\lambda^\star_\infty}{\Lambda'(\Lambda^{-1}(\hat\lambda^\star_\infty))} + \Lambda^{-1}(\hat\lambda^\star_\infty)\Bigg] - \hat\lambda^\star_\infty\xi_n - \kappa\delta_n + o(\epsilon_n)\Bigg], \tag{8}$$

where ξ_n is the corresponding delay cost. Using the fact that $\hat\lambda^\star_\infty$ satisfies the first-order condition for optimality, we obtain $\hat\lambda^\star_\infty/(\Lambda'(\Lambda^{-1}(\hat\lambda^\star_\infty))) + \Lambda^{-1}(\hat\lambda^\star_\infty) = \kappa$, and thus the optimization problem reduces to

$$n\Big[\min_{\epsilon_n, \delta_n \in \mathbb{R}:\ \delta_n > \epsilon_n} \hat\lambda^\star_\infty\xi_n + \kappa(\delta_n - \epsilon_n)\Big].$$

Applying Proposition 1, it suffices to restrict attention to cases where $\delta_n - \epsilon_n = \gamma n^{-r/(r+1)} + o(n^{-r/(r+1)})$. Furthermore, we have $\xi_n = \Gamma(r+1)\gamma^{-r}n^{-r/(r+1)} + o(n^{-r/(r+1)})$, and so we obtain the following optimization problem ignoring the lower-order terms in n:

$$n^{1/(r+1)}\Big[\min_{\gamma>0} \hat\lambda^\star_\infty h\Gamma(r+1)\gamma^{-r} + \kappa\gamma\Big]. \tag{9}$$

This immediately yields the following near-optimal prescriptions.

Proposition 2. *Fix an $\epsilon \in \mathbb{R}$ arbitrarily. Let $\gamma^\star = (hr\hat\lambda^\star_\infty\Gamma(r+1)/\kappa)^{1/(r+1)}$ denote the solution to* (9)*. Then, for $r \geq 1$, the prescription $(\hat\lambda^\star_\infty + \epsilon n^{-r/(r+1)}, \hat\lambda^\star_\infty + [\epsilon + \gamma^\star]n^{-r/(r+1)})$ is optimal up to a negligible tolerance. To be specific, it yields a profit that is within $o(n^{1/(r+1)})$ of the optimal solution to* (6)*.*

It is worth noting that there are infinitely many near-optimal prescriptions possible, one for each ϵ. All of these prescriptions lead to a negligible tolerance with respect to the optimal values. Although we have two degrees of freedom in choosing ϵ_n and δ_n, we need only one degree of freedom to achieve the optimal profit within a negligible tolerance; that is, price and capacity can be effectively traded off against each other, and we need only one of these two levers. This

observation has been made before in other settings; see, for example, Randhawa and Kumar (2008). However, unlike in Proposition 1, the tolerance with which the prescriptions are specified is higher than that of their corresponding profit levels.

3.2. Concave Delay Costs $(0 < r < 1)$

We now consider the case $\xi = h\mathbb{E}W^r$ for $0 < r < 1$, i.e., the case of strictly concave costs. When delay costs are concave, the FCFS policy no longer minimizes the expected delay costs. In fact, in a single-server system, if customers may be served in any order, the LCFS policy produces the lowest delay cost (cf. Li 1996). However, in many settings, because of fairness concerns, the FCFS may be the only policy that can be implemented. For this reason, as well as for continuity of exposition, we first analyze the solution under the FCFS policy and then consider the (nonpreemptive) LCFS policy.

The optimization problem for a fixed scheduling policy for the single-server configuration is given by

$$\begin{aligned} \max_{p,\mu\geq 0} \quad & pn\Lambda(p+\xi) - \kappa\mu \\ \text{s.t.} \quad & \mu > n\Lambda(p+\xi) \\ & \xi = h\mathbb{E}W^r, \end{aligned} \tag{10}$$

where ξ is calculated under the fixed scheduling rule. The following result characterizes the asymptotic behavior of the optimal solution under FCFS scheduling.

Proposition 3 (FCFS Policy). *For delay costs of the form $\xi = h\mathbb{E}W^r$ with $0 < r < 1$, for any solution to* (10) *under the FCFS policy, the system operates in $r/(r+1)$-heavy traffic.*

Proposition 3 says that the form of the result for the single-server configuration is identical to that for $r \geq 1$ derived in Proposition 1; that is, although the traffic is actually "lighter" in this regime, the functional form is identical.

Turning to the LCFS policy, which minimizes the delay costs, the following result establishes the optimal regime.

Proposition 4 (Optimal Policy: LCFS). *For delay costs of the form $\xi = h\mathbb{E}W^r$ with $0 < r < 1$, for any solution to* (10) *under the LCFS policy, the system operates in* 1/2*-heavy traffic.*

Table 2 Comparison of the Optimal Solution for Linear and Square-Root Delay Costs

	Linear delay costs		Square-root delay costs ($r = 1/2$)			
	FCFS		LCFS		FCFS	
n	Traffic intensity	Profit	Traffic intensity	Profit	Traffic intensity	Profit
1	0.54	0.52	0.67	0.8	0.62	0.7
10	0.79	15.72	0.86	15.14	0.79	13.8
100	0.92	201.6	0.95	192.7	0.89	180.2

Notes. For a smaller scale ($n = 1$), the system with square-root delay costs has a higher traffic intensity and profit. As the scale increases, the system with linear costs generates a higher profit. As expected, in the case of square-root delay costs, the LCFS policy outperforms the FCFS policy.

This result implies that under the LCFS policy, the optimal operating regime is in fact the conventional 1/2-heavy traffic. So, not only does the form of the delay cost affect the nature of heavy traffic, but also the scheduling rule. And the scheduling rule we chose here is not a pathological one. It is indeed optimal if minimizing delay costs is the objective.

Table 2 provides an illustration of the above results by comparing the optimal solution (which is computed as a numerical solution of (10)) for the case of linear and square-root delay costs ($r = 1/2$) for different system sizes. The same demand function and cost parameters are used as for Table 1. We obtain analogous results: for small system sizes ($n = 1$), the system with square-root delay costs dominates the one with linear delay costs, although, even for moderate scales ($n = 10$), linear delay costs generate higher profits.

3.3. General Delay Costs: Going Beyond Power Functions

Our aim here is to develop insights for the case of general delay cost functions of the form $\xi = h\mathbb{E}d(W)$. We will consider delay functions $d\colon \mathbb{R}_+ \to \mathbb{R}_+$ that are differentiable, increasing, and satisfy $d(0) = 0$, $d(x) > 0$ for $x > 0$, and $\int_0^\infty d(x)e^{-\sigma x}\,dx < \infty$ for some $\sigma > 0$. We will fix the scheduling policy as FCFS. The manager's optimization problem for the single-server configuration can be explicitly written as

$$\begin{aligned} \max_{p,\mu\geq 0} \quad & pn\Lambda(p+\xi) - \kappa\mu \\ \text{s.t.} \quad & \mu > n\Lambda(p+\xi) \\ & \xi = h\frac{n\Lambda(p+\xi)}{\mu}(\mu - n\Lambda(p+\xi)) \\ & \qquad \cdot \int_0^\infty d(x)e^{-(\mu - n\Lambda(p+\xi))x}\,dx. \end{aligned}$$

Our earlier analysis implies that the optimal regime is one where the delays are extremely small. This suggests that it is the behavior of the delay function in the vicinity of zero that is crucial in determining the asymptotic behavior. The following result formalizes this notion.

Proposition 5. *If $d(x)/x^r \to C \in (0,\infty)$ as $x \to 0$ for some $r > 0$, then under the optimal solution, the system operates in $r/(r+1)$-heavy traffic.*

This result implies that if $d(x) \approx x$ in the vicinity of zero, we obtain 1/2-heavy traffic. This immediately translates into a condition on the derivative of the delay function at zero.

Corollary 1. *Under the optimal solution, the system operates in 1/2-heavy traffic if and only if $0 < d'(0) < \infty$.*

To better understand the above results, we present a few examples. If $d(x) = x^r + x^t$ for some $t > r$, the behavior of the delay function in the vicinity of zero will be dominated by x^r, and hence the optimal operating regime will be $r/(r+1)$-heavy traffic. If $d(x) = \log(1+x)$, we have $d(x) \approx x$ for x near zero or, in other words $d'(0) = 1$, and we obtain the optimal regime as 1/2-heavy traffic. In the case that $d(x)/x^r \to 0$ or ∞ as $x \to 0$ for all $r > 0$, we can only determine that $1 - \rho_n^\star$ cannot significantly differ from $\Theta(n^{-\tilde{r}/(\tilde{r}+1)})$ for some $\tilde{r} > 0$.

This section was devoted to analyzing the firm's joint optimization problem on price and capacity. There might be cases where a firm may not have both these levers available to optimize. Our results extend to these settings as well, and this is the subject of the next two sections. In §4, we study the case where the capacity is exogenously fixed, and the system manager optimizes on the price. In §5, we consider the case of exogenous prices, where the single decision variable is the capacity.

4. Pricing with Exogenous Capacity Growth: Illustrating When Heavy Traffic May Not Be Achieved

In this section, our goal is to analyze how a firm may exploit scale when the capacity level is exogenously determined or, more precisely, when the ratio of capacity to market size is fixed. So, the only decision variable is the price. A by-product of this analysis will be the fact that heavy traffic need not always be realized in this setting. In fact, an elasticity property of the demand function will be essential for heavy traffic to be optimal. Interestingly, flexibility on capacity choice makes this issue of elasticity moot in the general case.

The system manager's objective in this setting is to maximize the revenue, given by $np\Lambda(p + h\mathbb{E}W^r)$, by setting the appropriate price. The system capacity, or the service rate, is not a decision variable, and scales with the market size as $n\hat{\mu}$, where $\hat{\mu} > 0$ is fixed. The manager's optimization problem can be stated as

$$\begin{aligned} \max_{p\geq 0} \quad & np\Lambda(p+\xi) \\ \text{s.t.} \quad & \hat{\mu} > \Lambda(p+\xi) \\ & \xi = h\frac{\Lambda(p+\xi)}{\hat{\mu}} n^{-r} \frac{\Gamma(r+1)}{(\hat{\mu}-\Lambda(p+\xi))^r}. \end{aligned} \tag{11}$$

As the capacity level is predetermined, the behavior of the system will depend on the properties of the demand function, for example, whether it is elastic or not. To better understand this, let us for now disregard any variability in the model. In other words, let us consider the following version of (11) that ignores delay costs:

$$\begin{aligned} n\Big[\max_{p\geq 0} \quad & p\Lambda(p)\Big] \\ \text{s.t.} \quad & \hat{\mu} \geq \Lambda(p). \end{aligned} \tag{12}$$

Note that this problem follows from (11) by setting $\xi = 0$. Before discussing solutions to this problem, we define demand elasticity. The elasticity of a demand function Λ at a price p is given by $e(p) = -(\partial\Lambda(p)/(\partial p))(p/(\Lambda(p)))$. A demand function is said to be elastic on an interval $[a,b]$ if $e(p) > 1$ on $[a,b]$. Elasticity of a demand function implies that the corresponding revenue increases as the price is lowered. For such a demand function, the optimal solution to (12) will set the price as low as possible, which will imply that the constraint $\hat{\mu} \geq \Lambda(p)$ will hold with equality. Thus, the optimal solution will be $p^\star = \Lambda^{-1}(\hat{\mu})$. Such a demand function leads the system to a heavily loaded condition. Now, consider the case when the demand function is not elastic. Here, it may be the case that the optimal solution to (12) is such that $\Lambda(p^\star) < \hat{\mu}$, and hence the system is loaded lightly.

This analysis suggests that when dealing with a pricing problem, the elasticity of the demand curve dictates the behavior of the system.

We now make this informal reasoning rigorous. As in §3, we will rewrite (11) as a problem of selecting the optimal arrival rate instead of the optimal price to obtain the following equivalent problem:

$$\max_{0\le\hat{\lambda}<\hat{\mu}} \Pi_n(\hat{\lambda}) \equiv n\left[\left(\Lambda^{-1}(\hat{\lambda}) - hn^{-r}\frac{\hat{\lambda}}{\hat{\mu}}\frac{\Gamma(r+1)}{(\hat{\mu}-\hat{\lambda})^r}\right)\hat{\lambda}\right]. \quad (13)$$

Let $\Pi_n^\star$ and $\hat{\lambda}_n^\star$ denote the optimal objective function and any optimizer of (13), respectively. Then, the optimal price $p_n^\star = \Lambda^{-1}(\hat{\lambda}_n^\star) - h(\hat{\lambda}_n^\star/\hat{\mu})n^{-r}(\Gamma(r+1)/(\hat{\mu}-\hat{\lambda}_n^\star)^r)$. The corresponding analog of (12) is given by

$$\max_{0\le\bar{\lambda}\le\hat{\mu}} \bar{\Pi}(\bar{\lambda}) \equiv \Lambda^{-1}(\bar{\lambda})\bar{\lambda}. \quad (14)$$

Let $\Pi_\infty^\star$ and $\hat{\lambda}_\infty^\star$ denote the optimal objective function and any optimizer of problem (14), respectively. Then, either $\hat{\lambda}_\infty^\star = \hat{\mu}$ or $\hat{\lambda}_\infty^\star < \hat{\mu}$. The former case corresponds to a critical match between capacity and arrival rate, whereas the latter corresponds to a capacity surplus. For an infinitely differentiable function $f\colon \mathbb{R}\to\mathbb{R}$, we use the notation $f^{(i)}$ to denote its ith derivative with the convention that $f^{(0)}(\cdot)\equiv 0$. Define $j \equiv \arg\max\{l\colon \bar{\Pi}^{(i)}(\hat{\lambda}_\infty^\star) = 0\,\forall\, i < l\}$. Assuming for convenience that Λ is infinitely differentiable, we obtain the following result. (See Lemma 1 in the appendix for the general version.)

Proposition 6. *For $r \ge 1$, when (14) has a unique solution, we have the following.*

1. *If $\hat{\lambda}_\infty^\star = \hat{\mu}$:*

(a) *Under the optimal solution, the system operates in $r/(r+j)$-heavy traffic.*

(b) *We have $\Pi_n^\star = n[\Pi_\infty^\star - \Theta(n^{-rj/(r+j)})]$.*

(c) *The prescription $\hat{\lambda}_n^p = \hat{\mu} - (h\Gamma(r+1)(j-1)!r\hat{\mu}/(-1)^{j-1}\bar{\Pi}^{(j)}(\hat{\mu}))^{1/(r+j)}n^{-r/(r+j)}$ has a negligible tolerance. In particular, the optimal solution $\hat{\lambda}_n^\star = \hat{\lambda}_n^p + o(n^{-r/(r+j)})$. Furthermore, $\Pi_n(\hat{\lambda}_n^p) = \Pi_n^\star + o(n^{1-(rj/(r+j))})$.*

2. *If $\hat{\lambda}_\infty^\star < \hat{\mu}$, it is not optimal to operate in heavy traffic.*

Thus, it is optimal to approach heavy traffic only if the demand function is elastic in the vicinity of $p^\star = \Lambda^{-1}(\hat{\mu})$. Note that when the system does approach heavy traffic, as the deterministic profit function $\hat{\lambda}\Lambda^{-1}(\hat{\lambda})$ gets "flatter" around its maxima $\hat{\lambda}_\infty^\star$ (as captured by an increase in the number of its derivatives being equal to zero at that point), $\rho_n^\star$ approaches heavy traffic more slowly, and the solution $\Pi_n^\star$ gets closer to the deterministic upper bound $n\Pi_\infty^\star$. Intuitively, this makes sense. When $\hat{\lambda}\Lambda^{-1}(\hat{\lambda})$ is nearly constant in the vicinity of the maxima, it allows one to choose smaller arrival rates to reduce the delay costs, while not affecting the revenue component of the objective function significantly.

We next consider the complementary case where the arrival rate is exogenously fixed and only the capacity can be chosen by the firm. This scenario is equivalent to the case where prices are fixed.

5. Capacity Sizing at Fixed Demand: Minimizing Delay Costs

In this section, we set aside pricing considerations by fixing the demand level exogenously, that is, we assume that customers arrive at a fixed rate $n\hat{\lambda}$. This case serves to complete the analysis for settings with only one decision variable. In the absence of pricing, the firm's objective is to select a capacity level to minimize the total delay costs experienced by the customers and the cost of capacity. The optimization problem is

$$\begin{aligned} \min_{\mu>n\hat{\lambda}} \quad & \Pi_n(\mu) \equiv n\hat{\lambda}\xi + \kappa\mu \\ \text{s.t.} \quad & \xi = h\frac{n\hat{\lambda}}{\mu}\frac{\Gamma(r+1)}{(\mu-n\hat{\lambda})^r}. \end{aligned} \quad (15)$$

Such a problem is studied for a many-server configuration in Borst et al. (2004). Using the first-order conditions that the optimal solutions must satisfy, we obtain the following characterization.

Proposition 7. *For $r \ge 1$:*

1. *Under any solution of (15), the system operates in $r/(r+1)$-heavy traffic.*

2. *The prescription $\mu_n^p = n[\hat{\lambda} + (hr\hat{\lambda}\Gamma(r+1)/\kappa)^{1/(r+1)} \cdot n^{-r/(r+1)}]$ has a negligible tolerance. In particular, the optimal capacity $\mu_n^\star = \mu_n^p + o(n^{1/(r+1)})$. Furthermore, $\Pi_n(\mu_n^p) = \Pi_n(\mu_n^\star) + o(n^{1/(r+1)})$.*

Part (a) of the result states that the optimal operating regime is $r/(r+1)$-heavy traffic. Part (b) of the

result characterizes a prescription that is near-optimal in the sense that the prescription as well as the cost generated by using the prescription are both within a negligible tolerance of the optimal values. This tolerance is of the order $o(n^{1/(r+1)})$.

As we will note in the next section, for strictly convex costs ($r > 1$), the single- and many-server configurations have identical solutions. Thus, Proposition 7 will apply to many-server configurations as well, and is consistent with the observations in Borst et al. (2004).

6. Many-Server Configuration: Asymptotic Equivalence to Single-Server Configuration for Strictly Convex Delay Costs ($r > 1$)

We now turn our attention to the many-server configuration. Here, the firm deploys a number of servers denoted by s, each working at a fixed rate μ, to maximize profits. We will focus on the joint pricing and capacity sizing problem; results for the other settings follow similarly. We use the same notation as before with the exception that κ now denotes the cost per server deployed. We fix $r > 1$, i.e., strictly convex delay costs in the set of delay costs we consider. The manager's optimization problem here is given by

$$\begin{aligned} \max_{p,\, s\geq 0,\, s\in\mathbb{Z}_+} \quad & pn\Lambda(p+\xi) - \kappa s \\ \text{s.t.} \quad & \mu s > n\Lambda(p+\xi): \\ & \xi = h\alpha(n\Lambda(p+\xi), s)\frac{\Gamma(r+1)}{(s\mu - n\Lambda(p+\xi))^r}, \end{aligned} \tag{16}$$

where

$$\alpha(n\Lambda(p+\xi), s) = \frac{(n\Lambda(p+\xi)/\mu)^s/(s!(1-\rho))}{(n\Lambda(p+\xi)/\mu)^s/(s!(1-\rho)) + \sum_{k=0}^{s-1}((n\Lambda(p+\xi)/\mu)^k/k!)}, \tag{17}$$

and $\rho = n\Lambda(p+\xi)/(s\mu) < 1$ is the traffic intensity. Note that to compute the delay cost, we use the fact that the steady-state delay distribution $\mathbb{P}(W > t) = \alpha(n\Lambda(p+\xi), s)e^{-(s\mu - n\Lambda(p+\xi))t}$, $t \geq 0$ (see, for example, Chan and Lin 2003, Equation (13)).

An equivalent formulation in terms of maximizing over the arrival rate and scaling the arrival rate and number of servers by n analogous to (6) is

$$n\left[\max_{\hat{\lambda}\geq 0,\, \hat{s}\in\{x:\ \mu x > \hat{\lambda},\, nx\in\mathbb{Z}_+\}} \left(\Lambda^{-1}(\hat{\lambda}) - h\alpha(n\hat{\lambda}, n\hat{s})n^{-r} \cdot \frac{\Gamma(r+1)}{(\hat{s}\mu - \hat{\lambda})^r}\right)\hat{\lambda} - \kappa\hat{s}\right]. \tag{18}$$

Let $\Pi_n^\star$ and $(\hat{\lambda}_n^\star, \hat{s}_n^\star)$ denote the optimal objective function and any optimizer of (18), and $\Pi_\infty^\star$ and $(\hat{\lambda}_\infty^\star, \hat{s}_\infty^\star)$ denote the solution and any optimizer of the idealized system $\max_{\bar{\lambda}\geq 0,\, \bar{s}\geq\bar{\lambda}/\mu} \Lambda^{-1}(\bar{\lambda})\bar{\lambda} - \kappa\bar{s}$, respectively. As before, we assume that $0 < \hat{\lambda}_\infty^\star < \infty$. Then, we obtain the following result.

Proposition 8. *For $r > 1$:*

1. *Under any solution of* (18), *the system operates in $r/(r+1)$-heavy traffic.*
2. *The optimal profit is given by*

$$\Pi_n^\star = n[\Pi_\infty^\star - \Theta(n^{-r/(r+1)})].$$

3. *Fix an $\epsilon \in \mathbb{R}$ arbitrarily and let*

$$\gamma_m^\star = (hr\hat{\lambda}_\infty^\star\Gamma(r+1)\mu/\kappa)^{1/(r+1)}.$$

Then, the prescription $\hat{\lambda}_\infty^\star + \epsilon n^{-r/(r+1)}$,

$$\frac{1}{\mu}\left[\hat{\lambda}_\infty^\star + [\epsilon + \gamma_m^\star]n^{-r/(r+1)}\right]$$

is optimal up to a negligible tolerance. To be specific, it yields a profit that is within $o(n^{1/(r+1)})$ of the optimal solution to (18).

This result is identical to that for the single-server configuration (Propositions 1 and 2). The equivalence between single- and many-server configurations in $r/(r+1)$-heavy traffic goes beyond that indicated by this result. A straightforward calculation yields that the distributions of the steady-state delay in these systems are in fact identical up to a negligible tolerance.

For the case $r = 1$, the prescription for the many-server configuration is indeed different from the single-server configuration. This analysis is similar to that in Maglaras and Zeevi (2003, §4). Though the authors consider a slightly different version of the many-server configuration, where the capacity is shared between customers if their number exceeds the number of servers, their results can easily be applied to our setting. For brevity, we omit details.

In conclusion, we would like to point out that this equivalence breaks down for strictly concave costs.

The optimal operating regime in the many-server configuration for these costs when operating under the FCFS policy is, as expected, lighter than 1/2-heavy traffic. However, this implies that the behavior of the multiserver queue approaches that of an infinite-server queue rapidly, making the economical regime one that is only slightly lighter than 1/2-heavy traffic. In fact, it is heavier that $(1/2-\epsilon)$-heavy traffic for all $\epsilon > 0$. Unfortunately, we cannot analyze the optimal operating regime under the LCFS policy because the multiserver system does not have, to the best of our knowledge, a closed-form expression for the delay distribution.

7. Discussion

This paper studies a profit maximization problem in a large system setting. In particular, the system manager selects the capacity level and the price to be offered to a stream of price and delay sensitive customers. This paper deals with two configurations: single server, where the capacity is manifested in the service rate, and many servers, where the capacity is deployed in terms of the number of servers, each working at a fixed rate. We demonstrate that the optimal solution leads the system into a heavy traffic regime with utilization near 100%. The rate at which the utilization approaches 100% (with respect to the system size) depends on the form of the delay cost. In particular, for strictly convex costs (that have a zero derivative at the origin), the utilization approaches 100% at a rate faster than the conventional $O(1/\sqrt{n})$, and hence places the system in what is referred to as an efficiency driven regime (see Gans et al. 2003 for a discussion of the efficiency driven regime). In this regime, the single- and many-server configurations behave identically, and thus there is an easier method of approximating solutions for the typically difficult many-server settings. In the case of concave delay costs, we obtain similar results for the FCFS rule. However, in a single-server setting for the delay minimizing LCFS policy, we obtain 1/2-heavy traffic as the optimal regime. Thus, the nature of heavy traffic realized depends on the scheduling policy in addition to the cost structure. This suggests caution should be exercised when constructing optimal scheduling policies using an "assumed" heavy traffic regime.

The goal of this paper is to study the impact of the delay cost structure on the firm's pricing and capacity investment strategies in light of increasing market size. To facilitate this analysis, and to make explicit this interaction, we have made some simplifying assumptions as follows:

1. *Exponential assumptions.* This paper restricts attention to exponential interarrival and service times for customers. These assumptions are primarily for simplicity. For the single-server configuration, if there is no knowledge of service times and customers are processed in an FCFS fashion, our results remain for the case of general renewal arrivals and general services because asymptotically the delay distribution retains a similar structure. The extension for the many-server configuration is not as straightforward. Though the results are robust to general interarrival distributions, the case of general service distributions cannot be handled, primarily due to the lack of an accurate estimate of the delay distribution in this setting. We direct the reader to Whitt (1993) for a discussion of the approximation methods for the $GI/G/m$ queue.

2. *Linear capacity costs.* This paper assumes that the capacity cost is linear in the amount of capacity deployed. Such a cost structure is commonly used in the literature. (An exception is Borst et al. (2004), which studies a cost minimization problem in a many-server configuration for convex capacity costs.) For general capacity costs, the structure of these costs will also impact the optimal operating regime, in addition to the delay costs. In particular, fixing the structure of the delay costs, if there are economies of scale in acquiring capacity, then additional units of capacity will be cheaper, and it will be optimal to invest in higher levels of capacity. Thus, the optimal regime will be lighter than that suggested by the analysis for linear costs. Analogously, for diseconomies of scale, the optimal regime will be heavier than the linear cost case. The exact form of the capacity costs can be used to explicitly characterize the optimal regime.

3. *Additive demand functions.* This paper considers customer demand functions of the form $\Lambda(p+\xi)$, i.e., the customer disutility is additive in price and delay cost. This choice permits a natural separation of the price and delay cost, and an easy characterization of near-optimal prescriptions. One can envisage using

general demand functions $\Lambda(p, \xi)$. As an illustration, consider the case $\Lambda(p, \xi) = (1-p)(1-\xi)$. This is a multiplicative demand function that can be rewritten as $\Lambda(p, \xi) = 1-p-(1-p)\xi$. Considering the case with capacity cost $\kappa < 1$, analogous to the additive demand case, we note that as the market size increases, the delays become negligible, and the firm's optimal price is bounded away from 1 (in fact, the idealized price ignoring delay costs equals $(1+\kappa)/2$). Thus, the optimal regime for this demand function will be the same as that for the additive demand function $\hat{\Lambda}(p, \xi) = 1-p-k\xi$ for some appropriate constant $k > 0$. In this example, we are able to reduce the demand function to the case of an additive function. This may not be possible in general. However, the optimal regime can still be computed using the approach followed in this paper. We leave a detailed analysis for future research.

4. *Exact service times are not known.* If the system manager has knowledge of the customers' exact service times, then one expects policies such as shortest remaining processing time (SRPT) to outperform FCFS for convex delay costs (see, for example, Schrage and Miller 1966). By employing such a policy, the delay costs in the system can be further reduced, and a bounding argument yields that the optimal regime will be heavier than $r/(r+1)$-heavy traffic and will generate higher profits compared to the FCFS case. However, in the case where the customers' service times are bounded above, the delay costs under SRPT are of the same order as that under FCFS in heavy traffic. This result is obvious in the deterministic case, and for the general case, it follows by applying Jensen's inequality to the result in Wierman (2007, Theorem 3.16) that proves this relation for the first moment of delays. Hence, in this setting, the optimal operating regime will remain $r/(r+1)$-heavy traffic.

5. *Focus on time in queue.* This paper considers the delay incurred by customers during their wait for service. One can envisage considering the sojourn time instead. For the single-server configuration, our results remain the same because, in the optimal regime, the time spent in service will be of a smaller order than the delay in queue. For the many-server configuration, the distribution of the time spent in service remains unchanged, and the delay in queue diminishes with scale. In this case, with convex delay costs, using a Taylor series argument, one recovers the optimality of conventional 1/2-heavy traffic.

Appendix. Proofs of Results

Proof of Proposition 1. We will use the notation $g_n(\lambda, \mu) \equiv (\Lambda^{-1}(\lambda) - h(\lambda/\mu)n^{-r}(\Gamma(r+1)/(\mu-\lambda)^r))\lambda - \kappa\mu$, and $\overline{\Pi}(\lambda) \equiv \Lambda^{-1}(\lambda)\lambda$. The optimal solution to (6), $(\hat{\lambda}_n^\star, \hat{\mu}_n^\star)$, satisfies the following necessary first-order conditions:

$$\overline{\Pi}'(\hat{\lambda}_n^\star) = hn^{-r}\frac{\hat{\lambda}_n^\star}{\hat{\mu}_n^\star}\frac{\Gamma(r+1)}{(\hat{\mu}_n^\star - \hat{\lambda}_n^\star)^r}\left(2 + r\frac{\hat{\lambda}_n^\star}{\hat{\mu}_n^\star - \hat{\lambda}_n^\star}\right), \tag{19}$$

$$\kappa = hn^{-r}\frac{\hat{\lambda}_n^{\star 2}}{\hat{\mu}_n^\star}\frac{\Gamma(r+1)}{(\hat{\mu}_n^\star - \hat{\lambda}_n^\star)^r}\left(\frac{r}{(\hat{\mu}_n^\star - \hat{\lambda}_n^\star)} + \frac{1}{\hat{\mu}_n^\star}\right). \tag{20}$$

We now use the fact that all cluster points of $(\hat{\lambda}_n^\star, \hat{\mu}_n^\star)$ must be a solution of (7). Thus, we can write $\hat{\lambda}_n^\star = \hat{\mu}_n^\star - \epsilon_n$, where $\epsilon_n \to 0$, and we can rewrite (20) as

$$\kappa = hn^{-r}\frac{\hat{\lambda}_n^{\star 2}}{\hat{\mu}_n^\star}\Gamma(r+1)\epsilon_n^{-r}\left(r\epsilon_n^{-1} + \frac{1}{\hat{\mu}_n^\star}\right).$$

Using the fact that all solutions to (7) are strictly positive and finite, we obtain $C_1 \le \liminf_{n\to\infty} n^{-r}\epsilon_n^{-(r+1)} \le \limsup_{n\to\infty} n^{-r}\epsilon_n^{-(r+1)} \le C_2$ for finite constants $C_1, C_2 > 0$, which implies that $\rho_n^\star = 1 - \Theta(n^{-r/(r+1)})$.

We now turn to the optimal objective function. Using the Taylor series expansion of $\overline{\Pi}(\lambda) = \Lambda^{-1}(\lambda)\lambda$ around $\overline{\Pi}(\hat{\lambda}_\infty^\star)$, we can write

$$\begin{aligned}\frac{\Pi_n^\star}{n} &\ge g_n(\hat{\lambda}_\infty^\star - n^{-r/(r+1)}, \hat{\lambda}_\infty^\star)\\ &= \Pi_\infty^\star - n^{-r/(r+1)}\overline{\Pi}'(\zeta_n) - C_1 n^{-r/(r+1)}\hat{\lambda}_\infty^\star,\end{aligned} \tag{21}$$

where $\zeta_n \in [\hat{\lambda}_\infty^\star - n^{-r/(r+1)}, \hat{\lambda}_\infty^\star]$ and $C_1 > 0$ is a constant. We also have

$$\begin{aligned}\frac{\Pi_n^\star}{n} &= \overline{\Pi}(\hat{\lambda}_n^\star) - \kappa\hat{\mu}_n^\star - hn^{-r}\frac{\hat{\lambda}_n^\star}{\hat{\mu}_n^\star}\frac{\Gamma(r+1)}{(\hat{\mu}_n^\star - \hat{\lambda}_n^\star)^r}\hat{\lambda}_n^\star\\ &\overset{(a)}{\le} \overline{\Pi}(\hat{\lambda}_n^\star) - \kappa\hat{\lambda}_n^\star - C_2 n^{-r/(r+1)}\hat{\lambda}_n^\star\\ &\le \Pi_\infty^\star - C_2 n^{-r/(r+1)}\hat{\lambda}_n^\star,\end{aligned} \tag{22}$$

where $C_2 > 0$ is a constant and (a) follows by noting that $\hat{\mu}_n^\star > \hat{\lambda}_n^\star$ and $\rho_n^\star = 1 - \Theta(n^{-r/(r+1)})$. Combining (21) and (22), the result follows. □

Proof of Proposition 3. This is identical to the proof of Proposition 1 and is omitted. □

Proof of Proposition 4. For a fixed arrival rate λ and service rate μ, applying Corollary 9.3 in Chapter III of Asmussen (2003), the steady-state delay distribution is given by

$$\mathbb{P}(W > t) = \rho - \sqrt{\rho}\int_0^t \frac{1}{y}e^{-(\lambda+\mu)y}I_1(2y\sqrt{\lambda\mu})\,dy, \tag{23}$$

where $I_1(x) = \sum_{k=0}^{\infty}(x/2)^{2k+1}/(k!(k+1)!)$ denotes the Bessel function. Using this distribution, we can compute

$$\mathbb{E}W^r = \frac{\lambda}{(\lambda+\mu)^{r+1}}\Gamma(r+1)_2F_1\left(\frac{r+1}{2}, \frac{r+2}{2}; 2; \frac{4\lambda\mu}{(\lambda+\mu)^2}\right), \quad (24)$$

where ${}_2F_1$ is Gauss's hypergeometric function given by ${}_2F_1(a,b;c;z) = \sum_{m=0}^{\infty}(a)_m(b)_m/(c)_m)(z^m/m!)$, where $(y)_m = \prod_{k=0}^{m-1}(y+k)$ for $m>0$ and $(y)_0 = 1$.

We now turn to the optimization problem (10). We use the notation $\overline{\Pi}(\lambda) = \Lambda^{-1}(\lambda)\lambda$ and

$$g_n(\lambda,\mu) = \left[\Lambda^{-1}(\lambda) - hn^{-r}\frac{\lambda}{(\lambda+\mu)^{r+1}}\Gamma(r+1)_2F_1\frac{r+1}{2}, \frac{r+2}{2}; 2; \frac{4\lambda\mu}{(\lambda+\mu)^2}\right]\lambda - \kappa\mu.$$

Then, (10) is equivalent to $\max_{\hat{\lambda},\hat{\mu}\geq 0,\hat{\lambda}<\hat{\mu}} ng_n(\hat{\lambda},\hat{\mu})$. Any optimizer $(\hat{\lambda}_n^\star, \hat{\mu}_n^\star)$ must satisfy the first-order condition $\partial g_n(\hat{\lambda}_n^\star, \hat{\mu}_n^\star)/(\partial\mu) = \kappa$. This is equivalent to

$$\begin{aligned}&2(\hat{\lambda}_n^\star + \hat{\mu}_n^\star)^2\,{}_2F_1\left(\frac{1+r}{2}, \frac{2+r}{2}; 2; \frac{4\hat{\lambda}_n^\star\hat{\mu}_n^\star}{(\hat{\lambda}_n^\star+\hat{\mu}_n^\star)^2}\right)\\ &\quad + (2+r)\hat{\lambda}_n^\star(\hat{\mu}_n^\star - \hat{\lambda}_n^\star)_2F_1\left(\frac{3+r}{2}, \frac{4+r}{2}; 3; \frac{4\hat{\lambda}_n^\star\hat{\mu}_n^\star}{(\hat{\lambda}_n^\star+\hat{\mu}_n^\star)^2}\right)\\ &= n^r\frac{2(\hat{\lambda}_n^\star+\hat{\mu}_n^\star)^{r+4}\kappa}{(1+r)\hat{\lambda}_n^\star\Gamma(1+r)}. \quad (25)\end{aligned}$$

As in the proof of Proposition 1, we can write $\hat{\lambda}_n^\star = \hat{\mu}_n^\star - \epsilon_n$, where $\epsilon_n \to 0$ as $n\to\infty$. Combining this with the fact that $\lim_{z\to 1}(1-z)_2^{a+b-c}F_1(a,b;c;z) = \Gamma(c)\Gamma(a+b-c)/(\Gamma(a)\Gamma(b))$ for $a+b>c$ and noting that $1 - 4\hat{\lambda}_n^\star\hat{\mu}_n^\star/(\hat{\lambda}_n^\star+\hat{\mu}_n^\star)^2 = \Theta(\epsilon_n^2)$, we obtain

$${}_2F_1\left(\frac{3+r}{2}, \frac{4+r}{2}; 3; \frac{4\hat{\lambda}_n^\star\hat{\mu}_n^\star}{(\hat{\lambda}_n^\star+\hat{\mu}_n^\star)^2}\right) = \Theta(\epsilon_n^{-(1+2r)}) \quad \text{for } 0<r<1,$$

$$\begin{aligned}&{}_2F_1\left(\frac{1+r}{2}, \frac{2+r}{2}; 2; \frac{4\hat{\lambda}_n^\star\hat{\mu}_n^\star}{(\hat{\lambda}_n^\star+\hat{\mu}_n^\star)^2}\right)\\ &\quad = \Theta(\epsilon_n^{(1-2r)}) \quad \text{for } \frac{1}{2}<r<1,\\ &\quad \leq \Theta(\epsilon_n^{-\delta}) \quad \text{for } 0\leq r\leq\frac{1}{2} \text{ and any } \delta>0.\end{aligned}$$

Using the above relations in (25), we obtain $\rho_n^\star = 1 - \Theta(n^{-(1/2)})$. □

Proof of Proposition 5. The proof is similar to that of Proposition 1, and we provide only a sketch. Using the fact that $d(x)/x^r \to C$ as $x\to 0$, we have that for any $\epsilon>0$, there exists $\delta>0$ such that $|d(x)/x^r - C| < \epsilon$ for $x<\delta$. We use this relation to derive bounds on the delay cost. For an $M/M/1$ queue with arrival rate $n\hat{\lambda}_n$ and service rate $n\hat{\mu}_n$ for some $0<\hat{\lambda}_n<\hat{\mu}_n$, we can write the delay cost as

$$\begin{aligned}\mathbb{E}d(W_n) &= \rho_n n\hat{\mu}_n(1-\rho_n)\int_0^\infty d(x)e^{-n\hat{\mu}_n(1-\rho_n)x}\,dx\\ &= \rho_n n\hat{\mu}_n(1-\rho_n)\int_0^\delta d(x)e^{-n\hat{\mu}_n(1-\rho_n)x}\,dx\\ &\quad + \rho_n n\hat{\mu}_n(1-\rho_n)\int_\delta^\infty d(x)e^{-n\hat{\mu}_n(1-\rho_n)x}\,dx\\ &\overset{(a)}{=} \rho_n\sigma\int_0^{\delta n\hat{\mu}_n(1-\rho_n)/\sigma} d\left(\frac{\sigma y}{n\hat{\mu}_n(1-\rho_n)}\right)e^{-\sigma y}\,dy\\ &\quad + \rho_n n\hat{\mu}_n(1-\rho_n)e^{-n\hat{\mu}_n(1-\rho_n)\delta}\\ &\quad \cdot\int_\delta^\infty d(x)e^{-n\hat{\mu}_n(1-\rho_n)(x-\delta)}\,dz, \quad (26)\end{aligned}$$

where (a) follows by the substitution $y = n\hat{\mu}_n(1-\rho_n)x/\sigma$.

Noting that $|d(x)/x^r - C| < \epsilon$ for $x<\delta$, $\int_0^\infty d(x)e^{-\sigma x}dx<\infty$, and that under the optimal solution we must have

$$\liminf_{n\to\infty} n\hat{\mu}_n(1-\rho_n) = \infty, \quad 0<\liminf_{n\to\infty}\hat{\mu}_n \leq \limsup_{n\to\infty}\hat{\mu}_n<\infty,$$

and $\lim_{n\to\infty}\rho_n = 1$, we obtain, for n sufficiently large, $\mathbb{E}d(W_n) = \Theta(n^{-r}(1-\rho_n)^{-r})$. The rest of the proof proceeds by a bounding argument analogous to that used in the proof of Proposition 1. □

Proof of Proposition 6. We will prove the following general result (that assumes Λ is only $k\geq 1$ times continuously differentiable) that subsumes Proposition 6.

Lemma 1. *When* (14) *has a unique solution, for*

$$j = \arg\max\{l\colon \overline{\Pi}^{(l)} \text{ exists}, \overline{\Pi}^{(i)}(\hat{\lambda}_\infty^\star) = 0\,\forall\, i<l\},$$

we have the following.

1. *Case* $\hat{\lambda}_\infty^\star = \hat{\mu}$:

(a) *If* $\overline{\Pi}^{(j)}(\hat{\lambda}_\infty^\star)\neq 0$, *we have* $\rho_n^\star = 1-\Theta(n^{-r/(r+j)})$ *and* $\Pi_n^\star = n\Pi_\infty^\star - \Theta(n^{1-(rj/(r+j))})$. *Furthermore, the prescription*

$$\hat{\lambda}_n^p = \hat{\mu} - \left(\frac{h\Gamma(r+1)(j-1)!r\hat{\mu}}{(-1)^{j-1}\overline{\Pi}^{(j)}(\hat{\mu})}\right)^{1/(r+j)} n^{-r/(r+j)}$$

has a negligible tolerance in the sense that $\hat{\lambda}_n^\star = \hat{\lambda}_n^p + o(n^{-r/(r+j)})$ *and* $\Pi_n(\hat{\lambda}_n^p) = \Pi_n^\star + o(n^{1-rj/(r+j)})$.

(b) *Otherwise, we have* $\rho_n^\star = 1 - n^{-r/(r+k)}u(n)$ *for some function* $u(n)\geq 0$ *such that* $u(n)\to\infty$ *and* $u(n) = o(n^{r/(r+k)})$, *and* $n(\Pi_\infty^\star - K_1n^{-rk/(r+k)}u(n)^{-r} - K_2n^{-rk/(r+k)}u(n)^kv(n)^{-1}) \leq \Pi_n^\star \leq n(\Pi_\infty^\star - K_3n^{-rk/(r+k)}u(n)^{-r})$ *for some function* $v(n)$ *such that* $v(n)\to\infty$ *and finite constants* K_1, K_2, K_3 *with* $K_1, K_3>0$.

2. *Case* $\hat{\lambda}_\infty^\star < \hat{\mu}$:

(a) *If* $\overline{\Pi}^{(j)}(\hat{\lambda}_\infty^\star)\neq 0$, *we have* $\rho_n^\star = \hat{\lambda}_\infty^\star/\hat{\mu} - \Theta(n^{-r/(j-1)})$ *and* $\Pi_n^\star = n\Pi_\infty^\star - \Theta(n^{1-r})$. *Furthermore, the prescription*

$$\hat{\lambda}_n^p = \hat{\lambda}_\infty^\star - \left(\frac{(j-1)!h\Gamma(r+1)\hat{\lambda}_\infty^\star(2+r(\hat{\lambda}_\infty^\star/(\hat{\mu}-\hat{\lambda}_\infty^\star)))}{\hat{\mu}(\hat{\mu}-\hat{\lambda}_\infty^\star)^r(-1)^{j-1}\overline{\Pi}^{(j)}(\hat{\lambda}_\infty^\star)}\right)^{1/(j-1)} n^{-r/(j-1)}$$

has a negligible tolerance in the sense that $\hat{\lambda}_n^\star = \hat{\lambda}_n^p + o(n^{-r/(j-1)})$ *and* $\Pi_n(\hat{\lambda}_n^p) = \Pi_n^\star + o(n^{1-r})$.

(b) *Otherwise, we have* $\rho_n^\star = (\hat\lambda_\infty^\star/\hat\mu) - n^{-r/(k-1)}u(n)$ *for some function* $u(n)$ *such that* $|u(n)| \to \infty$ *and* $u(n) = o(n^{r/(k-1)})$, $n(\Pi_\infty^\star - K_1 n^{-r} - K_2 n^{-rk/(k-1)}u(n)^k v(n)^{-1}) \le \Pi_n^\star \le n(\Pi_\infty^\star - K_3 n^{-r})$ *for some function* $v(n)$ *such that* $v(n) \to \infty$ *and finite constants* K_1, K_2, K_3 *with* $K_1, K_3 > 0$. □

Proof. Consider the case $\hat\lambda_\infty^\star = \hat\mu$. We first prove the results for the optimal traffic intensity before moving on to the optimal objective function. Note that the optimal solution $\hat\lambda_n^\star$ satisfies the following necessary first-order condition:

$$\overline\Pi^{(1)}(\hat\lambda_n^\star) - hn^{-r}\frac{\hat\lambda_n^\star}{\hat\mu}\frac{\Gamma(r+1)}{(\hat\mu-\hat\lambda_n^\star)^r}\left(2+r\frac{\hat\lambda_n^\star}{\hat\mu-\hat\lambda_n^\star}\right)=0. \tag{27}$$

It is easy to see that we must have $\hat\lambda_n^\star \to \hat\lambda_\infty^\star$. Thus, we can write $\hat\lambda_n^\star = \hat\lambda_\infty^\star - \epsilon_n = \hat\mu - \epsilon_n$, where $\epsilon_n > 0$ and $\epsilon_n \to 0$ as $n \to \infty$. Substituting this relation in (27), we obtain

$$\overline\Pi^{(1)}(\hat\mu-\epsilon_n) = hn^{-r}\left(1-\frac{\epsilon_n}{\hat\mu}\right)\Gamma(r+1)\epsilon_n^{-r}((2-r)+r\hat\mu\epsilon_n^{-1}). \tag{28}$$

Using the definition of j, we apply the Taylor series expansion to $\overline\Pi^{(1)}$ about $\hat\mu$ in (28) to obtain

$$\frac{\overline\Pi^{(j)}(\zeta_n)}{(j-1)!}(-\epsilon_n)^{j-1} = hn^{-r}\left(1-\frac{\epsilon_n}{\hat\mu}\right)\Gamma(r+1)\epsilon_n^{-r}((2-r)+r\hat\mu\epsilon_n^{-1}), \tag{29}$$

where $\zeta_n \in [\hat\mu-\epsilon_n, \hat\mu]$. Dividing both sides by ϵ_n^{j-1} and taking the limit as $n \to \infty$ on both sides, we obtain

$$\lim_{n\to\infty} n^{-r}\epsilon_n^{-(r+j)} = \frac{(-1)^{j-1}\overline\Pi^{(j)}(\hat\mu)}{h\Gamma(r+1)(j-1)!r\hat\mu}. \tag{30}$$

Thus, if $\overline\Pi^{(j)}(\hat\mu) \ne 0$, we obtain the optimal traffic intensity $\rho_n^\star = 1 - \Theta(n^{-r/(r+j)})$. If $\overline\Pi^{(j)}(\hat\mu) = 0$, we must have $j = k$, and thus $\lim_{n\to\infty} n^{-r}\epsilon_n^{-(r+k)} = 0$, which gives us $\rho_n^\star = 1 - n^{-r/(r+k)}u(n)$ for some function $u(n) \ge 0$ such that $u(n) \to \infty$ and $u(n) = o(n^{r/(r+k)})$.

We now turn to the optimal objective function. Using $\hat\lambda_n^\star = \hat\lambda_\infty^\star - \epsilon_n$, the following relation holds:

$$n\left(\overline\Pi(\hat\lambda_n^\star) - hn^{-r}\Gamma(r+1)\left(1-\frac{\epsilon_n}{\hat\mu}\right)\epsilon_n^{-r}\hat\lambda_n^\star\right) = \Pi_n^\star \le n\left(\Pi_\infty^\star - hn^{-r}\Gamma(r+1)\left(1-\frac{\epsilon_n}{\hat\mu}\right)\epsilon_n^{-r}\hat\lambda_n^\star\right). \tag{31}$$

The result now follows for each case by using the corresponding limiting relation for ϵ_n and writing out the Taylor expansion around $\overline\Pi(\hat\lambda_\infty^\star)$. The properties of the presented prescription follow from (30) and by performing a calculation analogous to that above for the objective function.

We now consider the case $\hat\lambda_\infty^\star < \hat\mu$. As before, we set $\hat\lambda_n^\star = \hat\lambda_\infty^\star - \epsilon_n$, where $\epsilon_n \ne 0$ and $\epsilon_n \to 0$ as $n \to \infty$. Note that ϵ_n can now take negative values as well. We obtain the following analog of (28):

$$\overline\Pi^{(1)}(\hat\lambda_\infty^\star - \epsilon_n) = hn^{-r}\frac{\hat\lambda_\infty^\star - \epsilon_n}{\hat\mu}\frac{\Gamma(r+1)}{(\hat\mu-\hat\lambda_\infty^\star+\epsilon_n)^r}\left(2+r\frac{\hat\lambda_\infty^\star-\epsilon_n}{\hat\mu-\hat\lambda_\infty^\star+\epsilon_n}\right).$$

Using the Taylor series expansion around $\hat\lambda_\infty^\star$, we obtain the following analog to (29):

$$\begin{aligned}\frac{\overline\Pi^{(j)}(\zeta_n)}{(j-1)!}(-\epsilon_n)^{j-1} &= h\frac{n^{-r}\Gamma(r+1)(\hat\lambda_\infty^\star-\epsilon_n)}{\hat\mu(\hat\mu-\hat\lambda_\infty^\star)^r}\left[1-\frac{r\epsilon_n}{\hat\mu-\hat\lambda_\infty^\star}+o(\epsilon_n)\right]\\ &\quad\times\left(2+r\frac{\hat\lambda_\infty^\star}{\hat\mu-\hat\lambda_\infty^\star}-\frac{r}{\hat\mu-\hat\lambda_\infty^\star}\left(1+\frac{\hat\lambda_\infty^\star}{\hat\mu-\hat\lambda_\infty^\star}\right)\epsilon_n+o(\epsilon_n)\right),\end{aligned}$$

where $\zeta_n \in [\hat\lambda_\infty^\star - \epsilon_n, \hat\lambda_\infty^\star]$ for $\epsilon_n > 0$ and $\zeta_n \in [\hat\lambda_\infty^\star, \hat\lambda_\infty^\star - \epsilon_n]$ for $\epsilon_n < 0$. Dividing both sides by ϵ_n^{j-1} and letting $n \to \infty$, we obtain

$$\lim_{n\to\infty} n^{-r}\epsilon_n^{-(j-1)} = \frac{\hat\mu(\hat\mu-\hat\lambda_\infty^\star)^r(-1)^{j-1}\overline\Pi^{(j)}(\hat\lambda_\infty^\star)}{(j-1)!h\Gamma(r+1)\hat\lambda_\infty^\star(2+r(\hat\lambda_\infty^\star/(\hat\mu-\hat\lambda_\infty^\star)))},$$

and the rest of the argument follows as before. □

Proof of Proposition 7. Let $\mu_n^\star$ denote a solution to (15). Then, it must satisfy the following the first-order optimality condition:

$$h\frac{n^2\hat\lambda^2\Gamma(r+1)}{\mu_n^\star(\mu_n^\star - n\hat\lambda)^r}\left(\frac{1}{\mu_n^\star}+\frac{r}{\mu_n^\star - n\hat\lambda}\right)=\kappa. \tag{32}$$

It is easy to see that we must have $\mu_n^\star/n \to \hat\lambda$. Thus, we can write $\mu_n^\star = n\hat\lambda + n\epsilon_n$, where $\epsilon_n \to 0$ as $n \to \infty$. Using this in the above equation, we obtain

$$h\hat\lambda\Gamma(r+1)\frac{\hat\lambda}{\hat\lambda+\epsilon_n}n^{-r}\epsilon_n^{-r}\left(\frac{1}{\hat\lambda+\epsilon_n}+\frac{r}{\epsilon_n}\right)=\kappa.$$

Thus, we must have $n^r\epsilon_n^{(r+1)} \to h\hat\lambda\Gamma(r+1)r/\kappa$ as $n \to \infty$, and we obtain $\mu_n^\star = n\hat\lambda + (h\hat\lambda\Gamma(r+1)r/\kappa)^{1/(r+1)}n^{1/(r+1)} + o(n^{1/(r+1)})$. The result for the cost follows by a straightforward calculation. □

Proof of Proposition 8. To prove this result one can use the first-order necessary conditions for optimality analogous to the proof of Proposition 1. However, this approach requires calculating derivatives of the function $\alpha(\cdot,\cdot)$, which is fairly tedious. Instead we use a simpler bounding approach to prove the result. It is easy to see that all cluster points of the sequence $(\lambda_n^\star, s_n^\star)$ must be solutions to $\max_{\lambda\ge0,\, s\ge(\lambda/\mu)} \overline\Pi(\lambda) - \kappa s$. Noting that any solution to $\max_{\lambda\ge0,\, s\ge(\lambda/\mu)} \overline\Pi(\lambda) - \kappa s$ must satisfy the first-order optimality condition $\overline\Pi'(\hat\lambda_\infty^\star) = \kappa/\mu > 0$, and using the continuity of $\overline\Pi'$, we have

$$\lim_{n\to\infty}\overline\Pi'(\lambda_n^\star) = \frac{\kappa}{\mu}. \tag{33}$$

Using this, analogous to (21), we obtain the bound

$$\frac{\Pi_n^\star}{n} \geq \Pi_\infty^\star - C_1 n^{-r/(r+1)} \tag{34}$$

for some finite constant $C_1 > 0$. We also have the bound

$$\begin{aligned}\frac{\Pi_n^\star}{n} &= \overline{\Pi}(\hat{\lambda}_n^\star) - \kappa \hat{s}_n^\star - h\alpha(n\hat{\lambda}_n^\star, n\hat{s}_n^\star) n^{-r} \frac{\Gamma(r+1)}{\hat{s}_n^\star \mu (1-\rho_n^\star)^r} \hat{\lambda}_n^\star \\ &= \overline{\Pi}(\hat{\lambda}_n^\star) - \kappa \frac{\hat{\lambda}_n^\star}{\mu} - \kappa \hat{s}_n^\star (1-\rho_n^\star) \\ &\quad - h\alpha(n\hat{\lambda}_n^\star, n\hat{s}_n^\star) n^{-r} \frac{\Gamma(r+1)}{\hat{s}_n^\star \mu (1-\rho_n^\star)^r} \hat{\lambda}_n^\star \\ &\leq \Pi_\infty^\star - C_2(1-\rho_n^\star) - C_3 \alpha(n\hat{\lambda}_n^\star, n\hat{s}_n^\star) n^{-r} \frac{1}{(1-\rho_n^\star)^r}, \end{aligned} \tag{35}$$

for some finite constants $C_2, C_3 > 0$, where we use the fact that $\liminf_{n\to\infty} \lambda_n^\star > 0$, $\hat{\lambda}_n^\star < \mu \hat{s}_n^\star < \infty$.

If along some sequence denoted by $\tilde{n}$ we have $\tilde{n}^{r/(r+1)} \cdot (1-\rho_{\tilde{n}}^\star) \to \infty$ as $\tilde{n} \to \infty$, then for large $\tilde{n}$ we must have $C_1 \tilde{n}^{-r/(r+1)} < C_2(1-\rho_{\tilde{n}}^\star)$ and the bound in (35) contradicts that in (34). Similarly, we cannot have $\tilde{n}^{r/(r+1)}(1-\rho_{\tilde{n}}^\star) \to 0$ as $\tilde{n} \to \infty$ because this would give us $\sqrt{\tilde{n}}(1-\rho_{\tilde{n}}^\star) \to 0$, which by the standard results of Halfin and Whitt (1981, p. 575) implies $\alpha(\tilde{n}\hat{\lambda}_{\tilde{n}}^\star, n\hat{s}_{\tilde{n}}^\star) \to 1$, and we again obtain a contradiction between (34) and (35). Thus, we must have $\rho_n^\star = 1 - \Theta(n^{-r/(r+1)})$ and $\Pi_n^\star = n[\Pi_\infty^\star - \Theta(n^{-r/(r+1)})]$.

Part (3) follows by a Taylor series argument analogous to that for the single server configuration in §3.1, along with the fact that for $r > 1$, we have $\sqrt{n}(1-\rho_n^\star) \to 0$, and thus $\alpha(n\hat{\lambda}_n^\star, n\hat{s}_n^\star) \to 1$. □

References

Armony, M. 2005. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems: Theory Appl.* **51** 287–329.

Asmussen, S. 2003. *Applied Probability and Queues*, 2nd ed. Springer-Verlag, New York.

Ata, B., T. L. Olsen. 2009. Near-optimal dynamic leadtime quotation and scheduling under convex-concave customer delay costs. *Oper. Res.* **57**(3) 753–768.

Borst, S., A. Mandelbaum, M. Reiman. (2004). Dimensioning large call centers. *Oper. Res.* **52**(1) 17–34.

Chan, W.-C., Y.-B. Lin. 2003. Waiting time distribution for the $M/M/m$ queue. *IEE Proc. Commun.* **150**(3) 159–162.

Dai, J. G., W. Lin. 2008. Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Ann. Appl. Probab.* **18**(6) 2239–2299.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial review and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4**(3) 208–227.

Gurvich, I., W. Whitt. 2008. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* **11**(2) 237–253.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems. International Series in Operations Research and Management Science*. Kluwer Academic Publishers, Norwell, MA.

Kingman, J. F. C. 1961. The single server queue in heavy traffic. *Proc. Cambridge Phil. Soc.* **57** 902–904.

Li, J. 1996. From FIFO to LIFO: A functional ordering of service delay via arrival discipline. *J. Appl. Probab.* **33**(2) 507–512.

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Sci.* **49** 1018–1038.

Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* **53** 242–262.

Mandelbaum, A., A. L. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* **52**(6) 836–855.

Naor, P. 1969. The regulation of queue sizes by levying tolls. *Econometrica* **37** 15–24.

Plambeck, E. L., A. R. Ward. 2006. Optimal control of a high-volume assemble-to-order system. *Math. Oper. Res.* **31** 453–477.

Randhawa, R. S., S. Kumar. 2008. Usage restriction and subscription services: Operational benefits with rational users. *Manufacturing Service Oper. Management* **10**(3) 429–447.

Schrage, L. E., L. Miller. 1966. The queue $M/G/1$ with the shortest remaining processing time discipline. *Oper. Res.* **14**(4) 670–684.

Stolyar, A. L. 2004. Maxweight scheduling in a generalized switch: State space collapse and equivalent workload minimization under complete resource pooling. *Ann. Appl. Probab.* **14**(1) 1–53.

Tezcan, T. 2008. Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Math. Oper. Res.* **33**(1) 51–90.

Van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Ann. Appl. Probab.* **5**(3) 808–833.

Whitt, W. 1993. Approximations for the $GI/G/m$ queue. *Production Oper. Managment* **2**(2) 114–161.

Whitt, W. 2003. How multiserver queues scale with growing congestion-dependent demand. *Oper. Res.* **51**(4) 531–542.

Wierman, A. 2007. Scheduling for today's computer systems: Bridging theory and practice. Ph.D. thesis, Carnegie Mellon University, Pittsburgh.