Pricing and Capacity Sizing for Systems with Shared Resources: Approximate Solutions and Scaling Relations

Constantinos Maglaras
Columbia University

Assaf Zeevi*
Columbia University

To appear: *Management Science*

Abstract

This paper considers pricing and capacity sizing decisions, in a single-class Markovian model motivated by communication and information services. The service provider is assumed to operate a finite set of processing resources that can be *shared* among users, however, this shared mode of operation results in a service rate degradation. Users, in turn, are sensitive to the delay implied by the potential degradation in service rate, and to the usage fee charged for accessing the system. We study the equilibrium behavior of such systems in the specific context of pricing and capacity sizing under revenue and social optimization objectives. Exact solutions to these problems can only be obtained via exhaustive simulations, in contrast, we pursue *approximate solutions* that exploit *large capacity asymptotics*. Economic considerations and natural scaling relations demonstrate that the optimal operational mode for the system is close to "heavy traffic." This, in turn, supports the derivation of simple approximate solutions to economic optimization problems, via asymptotic methods that completely alleviate the need for simulation. These approximations seem to be extremely accurate. The main insights that are gleaned in the analysis are the following: Congestion costs are "small," the optimal price admits a two-part decomposition, and the joint capacity sizing and pricing problem decouples and admits simple analytical solutions that are asymptotically optimal. All of the above phenomena are intimately related to *statistical economies of scale* that are an intrinsic part of these systems.

Keywords: shared resources, heavy traffic, equilibrium, pricing, many-server limits

1 Introduction and Overview

In recent years there has been an explosive growth in the usage of the Internet, and in particular in the scope of services made available under the auspices of the so-called "World-Wide-Web"

[*] Both authors are with Graduate School of Business, Columbia University, 3022 Broadway, New York, NY 10027. Email: `c.maglaras@columbia.edu` and `assaf@gsb.columbia.edu` The authors and the manuscript have both benefited from the careful reading and constructive comments of two anonymous referees and the Associate Editor.

(WWW). Examples include internet telephony, streaming audio, e-mail and information retrieval, to name but a few. In addition, many Internet visionaries foresee the future of personal computing, and in particular of hand-held appliances, as mainly providing an entry port to the WWW, with most common software being accessed on remote servers rather than locally on the desktop or hand-held device. The rapid development of this technology is a constant challenge for service providers, as economic objectives, viz, revenue management, are blended in with the structural properties of these systems.

This paper focuses on a class of systems that deliver communication and information services with the following four features: *large* capacity, an upper bound on the service rate that can be allotted to each user, resources that can be *shared* among the users, and *elastic* demand. They bare many similarities with traditional OR/OM models; demand is stochastic, customers are sensitive to both price and congestion effects, and there is a finite amount of resources (processing and storage capacity) that constrains the performance of the system. In contrast with traditional models, however, there is typically no physical inventory of goods and no queueing related phenomena in these systems; congestion effects arise from the sharing of resources by the users. Thus, delays are not due to jobs or users waiting to receive service, rather they arise from a rate of service that is lower than nominal. In particular, any number of users can simultaneously access the service, but when the combined demand exceeds a nominal load, the system's processing capacity gets divided among them, and the rate of service experienced by the users degrades. On the other hand, when the demand is less than the nominal load, resources cannot be "pooled" to provide better-than-nominal service to the users. Internet telephony, or bandwidth provision via cable-modems are examples of physical systems that share these features. [1]

Our formulation assumes that the pricing mechanism used by the service provider (SP) is a fixed "pay-per access" charge, i.e., users pay a static *usage fee* for accessing the system. This is easy to implement, transparent to the user, and provides a good starting point for the more complex problem of dynamic pricing. Taking into account the users' sensitivity to price and service level, and capacity related costs (initial investment, operational costs, and equipment devaluation), the service provider is faced with the problem of jointly selecting the system's capacity and a fixed pricing rule to optimize system profits (*monopoly pricing*) or total utility (*social pricing*).

This paper strives to contribute to the analysis and understanding of the above problems along three dimensions. The first dimension concerns *model formulation.* We propose a simple Markovian model for the dynamics of systems with finite capacity and resource sharing capabilities (queueing

[1] In many applications, there is a binding constraint on the rate at which the user can access these services (typically due to other capacity constraints, e.g., the maximum speed supporting the users' connection to the network), as well as intrinsic limitations on the extent to which resources can be pooled on the server end.

and inventory effects are absent in our formulation). This is essentially an $M/M/\infty$ model where the total available processing capacity shared by all busy servers is constrained and the capacity that can be dedicated to each individual user therefore is upper bounded. User utility depends on the fixed usage fee as well as the anticipated service degradation; these effects, in turn, are captured via a probabilistic choice model that determines user behavior and aggregate demand. Due to these dynamics, an intrinsic "feedback" between congestion and demand is present (an increase in the former reduces the latter and vice versa), thus, we invoke an *equilibrium* framework that determines the system's nominal operating point. Despite the Markovian nature of the model, the intrinsic equilibrium structure renders the exact analysis of the optimization objectives essentially intractable, thus, one has to resort to approximate procedures.

The second dimension highlights the relation and connections between *economic analysis*, and the *optimal operation* of the system. We essentially show that if the demand for services is elastic, then economic optimization of the class of systems described in the previous paragraph is achieved in an operational regime where high resource utilization and positive probability of congestion are both present. One informal interpretation of this result is that "heavy traffic is economically optimal," with the understanding that this statement holds under structural and economic assumptions on the model and demand process. This observation allows us, in turn, to port analysis tools that support straightforward approximate solutions to the original economic optimization problems. Moreover, the optimal operating regime provides a simple (but, perhaps, not immediately obvious) classification of the various scaling relations that prevail in such large capacity systems that are optimized subject to economic criteria. Most notably, users experience only minor service degradation effects in spite of the high resource utilization, which is the key driver behind the economic optimality of the heavy traffic regime. These effects are directly related to the nominal capacity in a manner that reveals *statistical economies of scale*, and are similar to what has been observed in other large stochastic service systems such as call centers and communication networks.

Finally, the third dimension concerns the *mode of analysis* and *solution methodology*. As mentioned previously, economic arguments lead to the optimality of heavy traffic, which, in turn, allows us to harness some simple analytical insights from that framework. In particular, we develop a tractable approximation to the original economic optimization problems that yield analytical characterizations of the equilibrium behavior, as well as the optimal pricing and capacity decisions. Simple "rules of thumb" for both pricing and capacity sizing follow from basic scaling relations. This mode of analysis culminates in a "recipe" that involves simple numerical calculations, thus, completely circumventing the need for tedious optimization rooted on exact analysis. The latter approach is used, however, to illustrate the accuracy of the results obtained via these approximations.

The outline of this paper is as follows. The introduction concludes with a short literature review. Section 2 presents the model and formulates the optimal pricing problem. Section 3 reviews some machinery, and analyzes the behavior of large capacity systems with shared resources. Section 4 studies the pricing problem under a revenue maximization objective, and Section 5 extends the analysis to the problem of jointly optimizing over system capacity and price. Section 6 discusses the social pricing counterpart. Finally, Section 7 offers some concluding remarks. Throughout, proofs are relegated to the Appendix.

The work we present here is most closely related to, and inspired by, two quite disparate strands of research. The first is the framework advocated in the paper of Mendelson (1985), and subsequently extended by Mendelson and Whang (1990); a closely related paper is by Stidham (1985). Our basic setup in which the service system is embedded in a micro-economic framework, and the optimal pricing problem is studied as an equilibrium model, follow these papers. Although queueing per se is not explicitly present in our problem, the second stream of research that inspires our analysis is rooted in queueing theory. In their seminal paper, Halfin and Whitt (1981) present a powerful framework for analyzing queueing systems with "many" servers. Whitt (1992) discusses how scaling relations that emerge in the so-called Halfin-Whitt regime give rise to simple "rules-of-thumb" for the design and dimensioning of large scale service systems (e.g., customer contact centers). Our approach blends the economic setup of Mendelson and Whang, in which the problem is phrased, with the analytical tools of Halfin and Whitt that are ultimately used to solve it.

The literature studying the use of pricing to manage the impact of externalities in congestion sensitive systems is extensive and appears to date back to Naor (1969); it has since been extended and generalized in various directions by numerous authors. In terms of blending an equilibrium analysis with insights from heavy traffic queueing theory, the recent paper by Van Mieghem (2000) is probably the most akin to ours, though the setup, objective and use of the asymptotic analysis are different. Revenue management in the context of Internet service provision was recently studied by Paschalidis and Tsitsiklis (2000) in a Markov decision process framework where customers are *only* assumed to be price sensitive. Although they focus on dynamic pricing, an important conclusion of their study is that static pricing rules can achieve near optimal performance. A similar insight was derived in Gallego and van Ryzin (1994) in the "classical" context of selling a set of goods within a finite time horizon. A recent review of the revenue management literature can be found in McGill and van Ryzin (1999).

The study of large capacity service systems in the spirit of Halfin and Whitt (1981) is currently an active area of research. Garnett *et al.* (2001) study approximate design rules for large call centers, while Borst *et al.* (2000) consider dimensioning of call centers; the latter is related to our capacity sizing problem of Section 5. The flavor of the analysis in our paper is close to that in the work of Armony and Maglaras (2001) and Whitt (2003). Both papers deal with multi-server queueing systems with congestion sensitive demand but without economic considerations, and study the system equilibrium behavior using asymptotic methods. The relation between economics and heavy traffic has also been highlighted in a recent paper by Harrison (2001), though the framework and asymptotic regime are both different than ours.

The model that we posit with capacity constraints and shared resources is closely related to the recent work of Das and Srikant (2000), who analyze the performance of a congested link in a packet switched network, also appealing to the limit theory of Halfin and Whitt. Finally, we refer to Courcoubetis *et al.* (2000), Gibbens and Kelly (1999) and the references therein for an overview of Internet-related pricing literature that is tangentially related to this paper.

2 Problem Formulation

The system model. External arrivals to the system, in the form of *connection requests*, are modelled as a homogenous Poisson process with rate Λ. Each user has an i.i.d. *service requirement*, which is assumed to be exponentially distributed with known mean $1/\mu$. Our stylized system model attempts to capture three important features of the physical system: finite capacity, no resource pooling when the system is under-utilized, and the capability to share processing resources among users. We model this as an $M/M/\infty$ system with capacity constrained to C units. Specifically, let $N(t)$ denote the number of users connected at time $t \geq 0$, then the processing rate experienced by the user is

$$\text{service rate} = \begin{cases} 1 & N(t) \leq C \\ \frac{C}{N(t)} & N(t) > C. \end{cases}$$

Without loss of generality, we take C to be an integer corresponding to the number of *nominal processing resources*, assuming that such a notion is applicable. The capacity limitations play a crucial role in our modelling framework. Each nominal processing resource can be *dedicated* to handle a separate user request, providing service at unit rate (measured, e.g., in Kbits/second), and resources cannot be pooled when $N(t) < C$. However, when the number of connection requests exceeds the nominal capacity C, processing resources are *shared* in an egalitarian way among users; the system is then said to be operating in a *congested state.* The system dynamics are equivalent to those of an $M/M/C$ queue, though queueing effects are absent in our conceptual formulation and quality of service is succinctly summarized in the *service rate* or *throughput rate* that is allocated to the user.

Economic structure, user choice model and system equilibrium. We assume that the service provider charges a fixed connection fee $\$p > 0$ when users access the system. Service degradation, due to *lost throughput* when the system is congested, is manifested in the form of *delay* experienced by the users. In communication and information services, users typically observe the rate of transmission, which indicates the degradation in service. In particular, it is reasonable (and, as we argue, not restrictive for our purposes) to assume that the rate of service serves as a proxy to assess delay in the following manner. The rate delivered to the user at time $t \geq 0$ can be written as $\min(C/N(t), 1)$, and the user conceives delay as the inverse of the latter. Associating the nominal service level with a unit delay (i.e., the situation where a dedicated resource is allocated to the user), we define the user's *conceived excess delay* [2] as

$$D(t) = \begin{cases} 0 & N(t) \leq C \\ \frac{N(t)}{C} - 1 & N(t) > C. \end{cases} \tag{1}$$

We assume that a cost of $\$q > 0$ is associated with each unit of excess delay that the user experiences.[3] Users have independent identically distributed (i.i.d.) *reservation prices* for the service, which are independent of the arrival process and service requirements. They join the system if their reservation price is greater or equal to the steady-state expected value of the "full price." Specifically, user valuations are denoted by $v \geq 0$ and are distributed according to a probability distribution P. We will assume that the cumulative distribution function F is continuously differentiable, the density f is everywhere positive on its support, and the first moment is finite.

As indicated previously, we will focus our attention on the *equilibrium steady-state* behavior of the system. An *equilibrium* roughly corresponds to a demand rate λ^* and a resulting congestion cost $q\mathbb{E}D^*$ such that both are time-independent and together satisfy the demand relationship[4]

$$\lambda^*(p) := \Lambda P\left(v > p + q\mathbb{E}D^*\right), \tag{2}$$

where Λ is the *maximal potential arrival rate* into the system. (The maximal potential arrival rate may also be interpreted as the effective "market size" for the applications offered by the service provider.) In particular, given the reservation price and the full cost of accessing the service, the risk-neutral user decides to join the system with probability $P(v \geq p + q\mathbb{E}D^*)$. In the sequel, a superscript '*' is used when various quantities are considered relative to the equilibrium distribution.

[2] In fact, in large capacity systems $D(t)/\mu$, due to a pathwise version of Little's law, is asymptotically a correct approximation to the *actual* excess delay. In our asymptotic analysis both formulations lead to the same analysis and results.

[3] Perhaps it is more realistic to assume that delay costs are convex increasing, as in Van Mieghem (2000), however, as we show in Section 3, systems with large capacity are characterized by "small" excess delay. Hence, convex delay costs reduce to the linear structure assumed above via a first order Taylor series approximation.

[4] To be precise, we say that for some price p the system admits a unique *equilibrium* if there exists a unique steady-state probability distribution for the process $(N(t) : t \geq 0)$, such the expected excess delay per-user when taken w.r.t. to this distribution, $\mathbb{E}D^*$, induces a time homogenous external arrival rate λ^* through (2) and λ^*, in turn, is consistent with the aforementioned steady-state distribution.

We note that users knowledge of their average excess delay may be due either to repeated visits to the system in which the excess delay is observed, or alternatively, we may assume that the service provider makes this information available to the users to facilitate their decision whether to join the system or not.[5] The reader should also note that the expected excess delay ($\mathbb{E}D^*$) depends on the price p though this dependence will be suppressed for notational clarity. The following result establishes the intuitive fact that the system *always* admits a unique equilibrium.[6]

Proposition 1 *For each capacity $C > 0$, and price $p > 0$ there exists a unique equilibrium.*

The user choice model determines the nature of the *demand function* (the two notions are essentially equivalent). That is, the choice model defines the exogenous arrival rate into the system according to (2). The following examples illustrate how different admissible user choice models induce different demand functions. With slight abuse of notation, set $\lambda(x) = \Lambda P(v > x)$.

Examples. (1) Linear demand: if v is distributed uniformly $U[0, \frac{\Lambda}{\alpha}]$, then $\lambda(x) = \Lambda - \alpha x$; (2) Exponential demand: if v is distributed exponentially with mean $\frac{1}{\alpha}$, then $\lambda(x) = \Lambda e^{-\alpha x}$; and, (3) Iso-elastic demand: if v is distributed Pareto with shape parameter $\alpha > 0$, then $\lambda(x) = \Lambda x^{-\alpha}$.

Now, the *elasticity* of a demand function at a price x is given by

$$\varepsilon(x) := -\frac{\partial \lambda(x)}{\partial x} \frac{x}{\lambda(x)}.$$

A demand function is said to be *elastic* over an interval $[a, b]$ if $\varepsilon(x) > 1$ for all $x \in [a, b]$. The key economic assumption that we impose is the following.

Assumption 1 The demand function $\lambda(x)$ is elastic over the set $\{x : 0 \le \lambda(x) \le C\mu\}$.

For example, in the first two of the three examples of demand functions above, demand is *elastic* for certain values of the price x and *inelastic* for others; in the third model elasticity is constant and equal to the shape parameter α. Intuitively, a demand function is elastic if a decrease in price (and increase in demand) result in an increase in the revenue rate $x\lambda(x)$. In terms of the probabilistic primitives, it is equivalent to saying that $xf(x)/\bar{F}(x) > 1$, where $\bar{F}(\cdot) = 1 - F(\cdot)$.

[5] It is more realistic to assume that state-dependent information is provided or observed by the users, however this information structure entails a transient analysis which is more involved and will not be pursued.

[6] The intuition behind this result is straightforward: for a fixed price, λ is strictly decreasing in the delay, while the delay $\mathbb{E}D$ is strictly increasing in ρ, and thus in λ. These observations indicate that a unique fixed point should exist. This point is the so-sought equilibrium.

Empirical studies of the demand for bandwidth seem to support the assumption of elastic demand; for example, Lanning *et al.* (2000) estimated the elasticity to be around 1.4-1.5.[7]

The service provider's objective. We assume that the service provider has at her disposal the four-tuple (q, Λ, μ, P) summarizing the congestion cost, the choice model and its parameters, and the service requirement. Suppose that capacity is fixed and given at C. With the price mechanism fixed, the service provider's objective is to choose the connection price (p) so as to optimize an economic objective. Specifically, we consider the goal of revenue maximization under the unique system equilibrium. (Section 6 discusses extensions to the problem of social welfare maximization.) This formulation assumes the service provider is operating in a monopolistic context. Under any pricing decision p, the system evolves as a continuous time Markov chain. Since every user that connects to the system pays \$$p$, and the probability of this happening in the next δt time units is roughly $\lambda \delta t$, we have that the *revenue rate* in equilibrium is

$$R(p) = p\lambda(p + q\mathbb{E}D^*). \tag{3}$$

We will consider the following two formulations:

1. *Optimal pricing:* Capacity of the system is fixed a-priori. A price p is sought that optimizes system revenues.
2. *Joint capacity sizing and pricing:* Capacity C together with a price p are sought that jointly optimize system revenues.

Both problems are considered also w.r.t. a social welfare objective in Section 6

Mode of analysis. As we have noted in the introduction, solving the aforementioned optimization problems is not straightforward. The reason is that the equilibrium framework requires us to solve the steady-state distribution of an $M/M/C$ model, and this cannot be done in closed form. (The distribution can only be expressed via a finite sum, where the number of terms as well as their magnitude depends on C, and the dependence on λ is quite complicated.) Perhaps more importantly, we believe that this mode of analysis only reveals some of the potentially interesting insights pertaining to the operation and economic value of this class of service systems. (For further discussion see Section 3.3.)

[7] While such detailed studies are not available for many other more recently introduced information services, we expect this assumption to hold true, especially in this early stage of their adoption. For example, informal inspection of the cable modem industry seems to indicate that, at least in its initial phase, prices have been dropping while demand has been increasing in a manner that overall revenues have also been increasing.

The other viable option is simulation-based optimization searching for the pair $(\lambda^*, \mathbb{E}D^*)$ that satisfies (2) for the given price p. This is quite cumbersome, as it involves generating a large number of sample paths to accurately estimate the steady-state quantities of interest. Furthermore, the equilibrium pairs $(\lambda^*, \mathbb{E}D^*)$ must be determined for a wide range of prices in order to search for the optimal usage fee p. The class of systems we are attempting to model here typically has large nominal capacity, thus the simulation-based approach tends to be computationally intensive. Just like the direct numerical approach, this method is not very revealing. Our approach will emphasize natural scaling relations resulting in "rules-of-thumb" that generate some insight at the expense of settling for approximate solutions. It will also illuminate an interesting connection between economic optimization and heavy traffic limit theory which is of interest in its own right.

3 Analysis of Large Capacity Systems

This section motivates and develops an approximating model for the original system when its capacity is "large"; these approximations highlight a series of fundamental scaling relations.

3.1 Background and underlying assumptions

To make the forthcoming results more transparent, we first consider the "physical" modes in which the system can operate. Recall that the steady-state probability of congestion is given by $\mathbb{P}\left(\text{congestion}\right) := \mathbb{P}\left(D^* \geq 0\right) = \mathbb{P}\left(N^* \geq C\right)$, where N^* denotes the number of steady state connections in a system with capacity C in equilibrium. Then, following Garnett *et al.* (2001), we consider the following three modes of operation.

- **Cost-driven regime:** the system is under-capacitated, users almost always experience service degradation. In this regime, $\mathbb{P}(\text{congestion}) \approx 1$.
- **"Rationalized" regime:** the system has *balanced* capacity, users experience degradation some of the times they use the system. In this regime, $\mathbb{P}(\text{congestion}) \approx \nu \in (0, 1)$.
- **Quality-driven regime:** the system is over-capacitated, users hardly ever experience service degradation. In this regime, $\mathbb{P}(\text{congestion}) \approx 0$.

From a purely operations standpoint, the *cost-driven* regime will tend to under-emphasize congestion effects, while the *quality-driven* one will tend to stress quality of service, at the expense of over-capacitating the system. The "rationalized" regime may potentially offer a good compromise in terms of trading-off congestion costs, capacity costs and utilization.

The use of the probability of congestion as a proxy for quality of service is quite common. As it was shown by Halfin and Whitt (1981) and later on argued in Whitt (1992) this is a natural and canonical measure of the system operational characteristics. In fact, in their 1981 paper, Halfin and Whitt showed that in an $M/M/C$ system, $\mathbb{P}(\text{congestion}) \to \nu \in (0,1)$ as C grows large if and only if the load in the system is of the form $\rho = 1 - \gamma/\sqrt{C}$; note that this is a statement for the "rationalized" regime. The parameters ν and γ are related through a one-to-one expression given later on, and the system behavior basically depends on this parameter. In addition, the "natural" scales emerging when capacity is large are of order $\sqrt{\text{capacity}}$. Specifically, queue lengths tend to be of that order, and as a result the waiting times will be of order $1/\sqrt{\text{capacity}}$ (this is the time required to clear the backlogs).

These observations have important design and operation ramifications for service networks. For example, systems that strive to offer a certain level of quality of service as measured by the probability of congestion ν, can do so while operating at increasing utilization rates as given by $1 - \gamma/\sqrt{C}$ as their capacity C grows larger. This demonstrates the *statistical economies of scale* enjoyed by large service systems, and explains, for example, the high utilization rates experienced in modern-day call centers that go hand-in-hand with good quality of service; see, e.g., Garnett *et al.* (2001). Another important manifestation of this fact is captured by the familiar *square-root staffing rule* that dictates that in order to handle a load of λ requests per unit time, the system should have about $\lambda + \gamma\sqrt{\lambda}$ servers (assuming $\mu = 1$).

The remainder of this section will characterize the asymptotic performance of our model operating in the "rationalized" regime, while its capacity and potential demand grow proportionally large. In particular, we will derive here an analogue of Proposition 2.1 of Halfin and Whitt (1981) that characterizes congestion in an $M/M/C$ queue. Our analysis extends the one in Halfin and Whitt (1981) along the following directions: congestion effects are manifested through sharing rather than queueing; the arrival rate into the system is governed by rational user behavior and captured via an equilibrium analysis; and, the system manager has the added capability of selecting a static price to affect demand. The key questions that will be addressed are the following: (a) how large is the congestion term given that a non-trivial fraction of all arriving users will suffer some level of congestion? (b) what is its effect on the equilibrium demand for the service? (c) what are (if any) the implications to the pricing rules that one needs to consider? Armed with these insights, Section 4 will show that if one considers the objective of revenue maximization, the "rationalized" regime turns out to be the *optimal* mode from a pure economics standpoint.

The validity and applicability of the theory we are about to expose hinges on one simple assumption: capacity and potential demand (market size) are large. In particular, we will let the potential demand Λ and the corresponding capacity C grow proportionally large according to

$C\mu = \kappa\Lambda$ for some appropriate $\kappa > 0$, while the user valuation distribution F is unaltered. [8] That is, the market potential Λ increases, the user characteristics (valuations) stay the same, and the system size grows proportionally to Λ in order to serve this larger market. Finally, in the absence of congestion effects $\bar{p}$ defined by

$$\bar{p} := \bar{F}^{-1}(C\mu/\Lambda) = \bar{F}^{-1}(\kappa) \tag{4}$$

is the price that matches demand to capacity; i.e., $\Lambda P(v \geq \bar{p}) = C\mu$.

3.2 Main results

We now turn to one of the main theoretical building blocks that are the key in deriving approximate solutions to the pricing and capacity sizing problems. In the sequel we will specify results in terms of "$C \to \infty$" highlighting that C is a system quantity. Implicit in this statement is that Λ is also growing such that $C\mu = \kappa\Lambda$. (In order to recover an approximation for the behavior of a given system of interest one should just "plug in" the appropriate values of C and κ into the derived expressions.) In the sequel we make explicit the dependence of the price on the system capacity using the notation $p(c)$.

Theorem 1 *Assume that potential demand Λ and the capacity C grow proportionally large such that $C\mu = \kappa\Lambda$ for some $\kappa > 0$. Then, for any usage fee $p(C) > 0$ such that*

$$\mathbb{P}\left(\text{congestion}\right) \;\to\; \nu \in (0,1) \qquad \text{as } C \to \infty, \tag{5}$$

it follows that,

(i) user demand scales as $\lambda^* = C\mu - \gamma^*\sqrt{C}\mu + o\left(\sqrt{C}\right)$ and system utilization scales as

$$\rho^* := \frac{\lambda^*}{C\mu} = 1 - \frac{\gamma^*}{\sqrt{C}} + o(1/\sqrt{C}),$$

where γ^* is uniquely defined in terms of ν.

(ii) The usage fee $p(C)$ must be of the form

$$p(C) = \bar{p} + \frac{\pi}{\sqrt{C}} + o(1/\sqrt{C}),$$

where $\bar{p}$ was defined in (4), and π is a function of γ^*.

[8] The assumption that the valuation distribution is unaltered as Λ changes may require that some other parameters of the distribution are scaled appropriately as functions of Λ. This depends on the specific form of that distribution. For example, in the linear demand model $v \sim U[0, \frac{\Lambda}{\alpha}]$, and thus in order to keep the distribution unaltered as Λ changes one needs to scale the price sensitivity α proportionally to Λ. For the other two examples, the valuation distributions only depend on the parameter α and thus requires no change.

(iii) System congestion scales as

$$\mathbb{E}[\text{delay per user}] \;=\; \mathbb{E}D^* \;=\; \frac{d}{\sqrt{C}} + o(1/\sqrt{C}),$$

where d is an explicit function of γ^*.

Hereafter we use the notation $f(x) = o(g(x))$ if and only if $f(x)/g(x) \to 0$, and $f(x) \approx g(x)$ to mean $f(x) = g(x) + o(\sqrt{x})$ or $f(x) = g(x) + o(1/\sqrt{x})$ as $x \to \infty$, depending on the context, and where no confusion arises. The theorem demonstrates that in the "rationalized" regime, the system operates "close" to full utilization according to $1 - \rho^*(C) \approx \gamma/\sqrt{C}$, and congestion is small, roughly $d/\sqrt{C}$. The latter highlights the *statistical economies of scale* that such systems enjoy, and the fact that the associated congestion becomes a second order phenomenon. This specifies the appropriate regime for asymptotic analysis. In addition, the price control must admit the decomposition $p \approx \bar{p} + \pi/\sqrt{C}$, where $\bar{p}$ is explicitly identified in terms of exogenous problem parameters (Λ, μ) and the capacity (C). Note that due to the assumption that potential demand and capacity scale proportionally, $\bar{p}$ is indeed a constant independent of the system size. This implies that the pricing problems reduce to selecting the optimal value of π that will optimally balance the lost revenues with the congestion costs.

Discussion. We turn to several comments on the content of the theorem, and the various quantities appearing there.

1. Converse implications. The theorem is in fact an *if and only if* result. That is, the assumption that the probability of congestion has a non-degenerate limit and each of the conditions (i)-(iii) are all equivalent. We do not spell this out, however the proof of the theorem should suggest why this is true.

2. Parameter relations. The second order parameters (π, γ, d) are all uniquely determined by ν, the limiting congestion probability. For our purposes, it will be more useful to consider π as the free variable. In fact, since our study focuses on pricing objectives and price is the intrinsic design variable, second order "congestion related effects" (ν, γ, d) should really be considered subordinate to π. We will pursue the analysis assuming that this shift in focus has been made. To explain the nature of the infinitesimal parameters, first note that ν is the limit of $\mathbb{P}(N^* \geq C)$ as $C \to \infty$, and is given explicitly in (2.3) of Halfin and Whitt (1981)

$$\nu = \frac{\phi(\gamma)}{\gamma\Phi(\gamma) + \phi(\gamma)}, \tag{6}$$

where $\phi(x)$ and $\Phi(x)$ denote the density and cumulative distribution function of a standard normal random variable. Note that $\nu \in (0, 1)$ implies that $0 < \gamma < \infty$. Given the one-to-one relationship in (6), we will henceforth omit further reference to ν, and focus attention on γ. The following gives relations between γ, d and π (see the Appendix for details)

$$\gamma = (\pi + qd(\gamma))\,\frac{f(\bar{p})}{\bar{F}(\bar{p})} \qquad \text{where} \qquad d(\gamma) = \frac{\phi(\gamma)}{\gamma(\gamma\Phi(\gamma) + \phi(\gamma))}. \tag{7}$$

The expression for $d(\gamma)$ can be easily derived by noting that $d := \lim_{C\to\infty} \sqrt{C}\mathbb{E}D^*$. Plugging in the expression for the expected excess delay given in (9), namely, $\sqrt{C}\mathbb{E}D^* = \rho\mathbb{P}(N^* \geq C)/(\sqrt{C}(1-\rho))$, noting that $\sqrt{C}(1-\rho) \to \gamma$, $\mathbb{P}(N^* \geq C) \to \nu$, and $\rho \to 1$ as $C \to \infty$, we get the stated relation using the expression for ν given in (6).

It remains to establish the uniqueness of γ (its unique value was denoted by γ^* in the statement of the theorem). This is resolved in the next proposition, where (8) follows from (7).

Proposition 2 *Assume (5) in Theorem 1 holds and fix any value of $\pi \in \mathbb{R}$. Then, $\gamma^*(\pi) > 0$ is uniquely defined as the solution of*

$$\frac{1}{q}\left(\frac{\bar{F}(\bar{p})}{f(\bar{p})}\gamma - \pi\right) = \frac{\phi(\gamma)}{\gamma(\gamma\Phi(\gamma) + \phi(\gamma))}. \tag{8}$$

3. System equilibrium calculations. Note that the system equilibrium is now characterized via the second order arrival rate γ^*, which in turn can be obtained through a simple numerical computation. As discussed previously, γ^* is the key quantity that is used to derive approximations for the equilibrium demand rate $\lambda^* \approx C\mu - \gamma^*\sqrt{C}\mu$, and the equilibrium congestion cost $q\mathbb{E}D^* \approx qd(\gamma^*)/\sqrt{C}$. These simple "rules-of-thumb" will facilitate the analysis of the optimization problems stated in Section 2, thus, allowing us to circumvent the difficulties involved in direct calculations of the equilibrium, or a simulation-based approach.

3.3 Comments on exact analysis and complete resource pooling

Exact analysis. As mentioned in the Introduction and Section 2, one can also undertake an exact analysis of the $M/M/C$ model. To illustrate the main ideas of this approach and contrast it with the approximate analysis pursued above, consider the profit maximization objective described in (3). Given the definition of the excess delay, $D = (N/C - 1)^+$, a straightforward calculation that follows from Section 1 in Halfin and Whitt (1981) yields that when the traffic intensity $\rho := \lambda^*/C\mu < 1$ the steady state congestion term is given by

$$\mathbb{E}D^* = \frac{\rho\nu(\rho, C)}{C(1-\rho)}, \tag{9}$$

where $\nu(\rho, C) = \mathbb{P}(N^* \geq C)$ is the probability that an arriving user finds the system in the congested state, explicitly given in equation (1.2) of Halfin and Whitt (1981). Taking derivatives with respect to the price \$$p$ at the equilibrium point in (3), the first order condition is given by

$$\frac{\partial R(p)}{\partial p} = \bar{F}(p + q\mathbb{E}[D^*(p)]) - pf(p + q\mathbb{E}[D^*(p)]) \left(1 + q\frac{\partial \mathbb{E}[D^*(p)]}{\partial p}\right) = 0 \ ,$$

where we have made explicit the dependence of the delay on the price p. After lengthy yet straightforward calculations this can be reduced to[9]

$$p = \frac{\bar{F}(p + q\mathbb{E}D^*)}{f(pq\mathbb{E}D^*)} + \frac{q\rho\nu(\rho, C)}{C(1-\rho)^2}\left(1 + C(1-\rho)^2 + \rho(1 - \nu(\rho, C))\right).$$

To solve this equation one needs to first be able to evaluate what is the equilibrium ρ for each choice of \$$p$, and then find the unique optimizer p^{rm}. Although this approach seems obvious, it is not straightforward to pursue. In Section 4 we undertake this analysis using the large capacity asymptotics that were discussed above. This approach is revealing in several regards.

An alternative model with complete resource pooling. We briefly discuss an alternative mode of analysis in which resources may be pooled not only when the system is congested. That is, unlike the model formulated in this paper, this alternative system always devotes its full capacity (C) to process user service requirements. The underlying Markov chain in this case is identical to that of an $M/M/1$ system, where the server works at rate C instead of unit rate. Our interest, as before, is focused on large capacity systems with high traffic flows. We therefore assume that the demand rate scales as $\lambda(C) = \lambda C$ while the average service requirement is, as before, held fixed at $1/\mu$.

Let W denote the total time a user spends in the system in steady-state. Then, for an $M/M/1$ system Little's law yields that $\mathbb{E}W = 1/(C\mu - C\lambda)$. Now, if we fix the utilization $\rho = \lambda/\mu$, then the expected time in the system decays like $1/C$ as C grows large. In the so-called "rationalized" regime, i.e., where the arrival rate scales like $\lambda(C) = C\mu - \gamma\sqrt{C}\mu$ as C grows large, the expected time in the system will scale like $1/\sqrt{C}$. These scaling relations are consistent with the ones present in an $M/M/C$ system: if $\rho < 1$, then the congestion experienced due to resource sharing decays as $1/C$ as C grows large; if the system is assumed to operating in the "rationalized" regime, the congestion scales like $1/\sqrt{C}$. Since the congestion delay seems to drive the user choice dynamics, it would appear that the simpler $M/M/1$ model should give rise to the same insights derived on the basis of the $M/M/C$ model (that does not support complete resource pooling). A closer inspection, however, reveals that the two models are actually quite different.

In the sequel, we focus on the rationalized regime (the following section will establish its economic optimality). The key step is to parse the expected time in the system $\mathbb{E}W$ into two compo-

[9] The authors would like to thank one of the referees for pointing out this explicit derivation.

nents: the actual service time, and the excess delay. Then, for the $M/M/1$ model the congestion cost is of order $1/\sqrt{C}$, while the service time is of order $1/C\mu$ and the total number of jobs in the system is of order $\sqrt{C}$. In sharp contrast, in the $M/M/C$ system the congestion term is of order $1/\sqrt{C}$, while the time in service is of order $1/\mu$ and the total number of jobs in the system is roughly C, plus or minus fluctuations of order $\sqrt{C}$. Thus, the models give rise to fundamentally different dynamics along the following two components: (i) the number of users in the system; and, (ii) the relative time scales in the system. In particular, in the $M/M/1$ system, congestion ($\approx 1/\sqrt{C}$) $\gg$ actual time in service ($\approx 1/C\mu$). In contrast, in the $M/M/C$ system, congestion ($\approx 1/\sqrt{C}$) $\ll$ actual time in service ($\approx 1/\mu$).

The above differences have an immediate bearing on economic analysis. To this end, note that the probability of a user joining the system is $\mathbb{P}(v \geq p + q \times (\text{congestion}))$. While this expression does not include the time in service, this is implicitly assumed to be accounted for in the valuation v. In other words, one could rewrite the choice probability as $\mathbb{P}(v' \geq p_C + \frac{q}{\mu} + q \times (\text{congestion}))$, while in the $M/M/1$ system the corresponding choice probability would be $\mathbb{P}(v' \geq p_1 + \frac{q}{C\mu} + q \times (\text{congestion}))$. Since the congestion terms are of order $1/\sqrt{C}$ in both cases and $\Lambda\mathbb{P}(\text{joining the system}) \approx C\mu$, it must be that $p_C + \dfrac{q}{\mu} \approx p_1 + \dfrac{q}{C\mu}$. As C grows large, the first order price term in the $M/M/1$ model, denoted by $\bar{p}_1$, will be such that $\bar{p}_1 = \bar{p} + \frac{q}{\mu}$. In particular, this implies that the economic decisions extracted on the basis of the two models will differ significantly, emphasizing the need to carefully match the "right" model to the application of interest. (We note in passing that systems that motivate the study in this paper are probably not adequately modeled using the complete resource pooling assumption.)

4 Revenue Maximization

The previous section established that the "rationalized" regime is a reasonable and appealing mode of operation for large capacity systems, insofar as it balances quality and efficiency considerations. This section builds on this observation, and blends it into the economic objective of the system manager, namely, revenue maximization. To that end, the main result of the section asserts that under the assumption that demand is elastic, the "rationalized" regime is the optimal mode to operate the system from purely economic considerations. Consequently, the insights of Theorem 1 can be applied towards solving the pricing problem.

Economic optimality of the "rationalized" regime. First, we show that as capacity grows, revenue maximization implies that the system approaches the heavy traffic regime. Second, we show that revenue maximization implies the price decomposition asserted in Theorem 1, and consequently the "rationalized" regime is economically optimal. The main assumption underlying our analysis is that capacity is "scarce," and this is spelled out by assuming that demand is *elastic* (i.e., $\varepsilon > 1$) in the feasible range $\lambda \in [0, C\mu)$.

Informally, the elasticity assumption implies that the service provider is always "better off" (as measured by her monopolistic objective) inducing a higher equilibrium arrival rate. That is, in the absence of congestion costs, she should set a price that induces the maximum arrival rate sustainable by the system, given by $C\mu$. This, of course, puts the system in heavy traffic and results in significant congestion costs, thus reducing the equilibrium arrival rate. The relevant question is by how much?

Extending the revenue rate definition in (3), we set, for each $C > 0$, $R(p, C) := p\Lambda P(v \geq p + q\mathbb{E}D^*)$, and denote the optimal price $p^{rm}(C) \in \operatorname{argmax}_p R(p, C)$. The associated equilibrium traffic intensity is denoted by $\rho^{rm}(C)$, making the dependence of these quantities on C explicit (in the sequel this may be omitted when no confusion arises). The following proposition asserts that under demand elasticity, profit maximization implies that large capacity systems operate in heavy-traffic.

Proposition 3 *Let Assumption 1 hold and assume in addition that capacity scales proportionally to potential demand. Then, $\rho^{rm}(C) \to 1$ as $C \to \infty$.*

The intuition behind this result is fairly straightforward. First, Theorem 1 has already established that as the system approaches heavy traffic along the "rationalized" regime ($\rho(C) \approx 1 - \gamma/\sqrt{C}$), congestion costs behave like $O(1/\sqrt{C})$ and become negligible. While this does not imply that the system manager would choose to operate the system in this regime, it does establish the feasibility of sustaining high utilization and good quality of service, simultaneously. The demand elasticity implies that if congestion is negligible, the service provider will price to induce a high rate of demand, since this results in increased revenues. [In a recent paper, Harrison (2001) considers an example of a queueing system that is driven by its economic and physical structure to operate in heavy-traffic; the set up and analysis there are quite different from the one we pursue here.]

Proposition 3 suggests that the optimal price ought to be close to $\bar{p}$, which is the static price that "places" the system in heavy traffic. It says nothing, however, about the actual structure of the optimal price, the associated level of congestion, how close is the system to the heavy traffic regime, and what are the revenues realized by the optimized system. The next theorem addresses all of these questions.

Theorem 2 *Let Assumption 1 hold and assume in addition that capacity scales proportionally to potential demand. Then, for large C, the optimal price is given by*

$$p^{rm}(C) = \bar{p} + \frac{\pi^{rm}}{\sqrt{C}} + o(1/\sqrt{C}),$$

where $\bar{p}$ is the price that places the system in heavy traffic given in (4), and π^{rm} is the defined via the following optimization problem

$$\pi^{rm} = \operatorname*{argmin}_{\pi \in \mathbb{R}} \; \{\bar{p}\gamma^*(\pi) - \pi\} \tag{10}$$

where $\gamma^*(\pi)$ is defined for each $\pi \in \mathbb{R}$ as the unique solution of (8).

Given (10), the first order condition that characterizes the revenue maximizing price is

$$\bar{p} = \frac{\bar{F}(\bar{p})}{f(\bar{p})} - q \left.\frac{\partial d(\gamma)}{\partial \gamma}\right|_{\gamma^*} . \tag{11}$$

This equation is used to define the optimal second order social price π^{rm}, taking as given the definition of $\bar{p}$ (the heavy traffic inducing price), and the notion of the equilibrium $\gamma^*(\pi)$.

Discussion and ramifications. One key insight that follows from Proposition 3 and Theorem 2 is that economically optimized systems with large capacity, shared resources, and elastic demand operate under high nominal resource utilization rates; i.e., they give rise to heavy traffic as the nominal operating point. It is important to note that although the statement of Theorem 2 bares resemblance to Theorem 1, it is not derived from it. However, the converse implications of Theorem 1 (see Remark 1) establish that the structure of the optimal price in itself implies that the system operates in the "rationalized" regime. Consequently, the equilibrium traffic intensity is given by $\rho^{rm}(C) = 1 - \gamma^{rm}/\sqrt{C} + o(1/\sqrt{C})$, where $\gamma^{rm} := \gamma^*(\pi^{rm})$ satisfies (8), or equivalently, the equilibrium demand is given by $C\mu - \gamma^{rm}\sqrt{C}\mu + o(\sqrt{C})$. These expressions describe the precise way by which the system should approach heavy traffic in order to optimally balance revenues with congestion costs. Moreover, the mode of operation for the "optimized" system is such that the probability of congestion is moderate, i.e., congestion is not completely avoided, nor is it the standard fare.

Theorem 2 specifies the form of the optimal price, and a simple numerical optimization routine that computes it to within very high accuracy. Specifically, this Theorem asserts that we can restrict attention to the "rationalized" regime, and to pricing rules of the form $p = \bar{p} + \pi/\sqrt{C}$. Let us first give some intuition pertaining to the derivation of the optimal price. Using the scaling relations of Theorem 1, the revenue rate can be expressed as:

$$\begin{aligned} R(p,C) &= \lambda^*(p+q\mathbb{E}D^*)p \\ &= \left(C\mu - \gamma^*(\pi)\sqrt{C}\mu + o(\sqrt{C})\right)\left(\bar{p} + \frac{\pi}{\sqrt{C}} + o(1/\sqrt{C})\right) \\ &\approx C\mu\bar{p} - \sqrt{C}\mu\left(\bar{p}\gamma^*(\pi) - \pi\right), \end{aligned}$$

where the first equality is a restatement of Theorem 1 parts (i) and (ii). Hence, as capacity grows large, the problem of selecting the optimal static price to maximize revenues reduces to the problem of choosing $\pi \in \mathbb{R}$ to minimize the second order term of lost revenues above, which is exactly (10) in the statement of the proposition. The tradeoff in (10) is evident: as π increases we have that $\lambda^*(p)$ decreases and thus $\gamma^*(\pi)$ increases (as a consequence of Theorem 1). Revenue maximization essentially reduces to a second order analysis that answers the following question: "how far from heavy traffic should the system be in order to optimally balance lost revenues and congestion costs?"

Algorithmically, we can compute the optimal price as follows. Taking as input the endogenous variable, i.e., fixed capacity (C), and the exogenous parameters of the problem, viz, potential demand (Λ), user choice model (F), and user mean service requirement ($1/\mu$), we proceed as follows. First, set the parameter $\bar{p}$ that appears in the limit problem in terms of original problem data as in (4), viz, $\bar{p} = \bar{F}^{-1}(\kappa)$, where $\kappa = C\mu/\Lambda$. Second, compute the equilibrium $\gamma^*(\pi)$ using (8) for each π in a fine grid. Finally, evaluate $\bar{p}\gamma^*(\pi) - \pi$ for each value of π in the grid, and seek the minimum value. The system revenues can be computed using the expansions for $R(p,C)$ given above.

The closed form expressions derived thus far can be used to check analytically or numerically the sensitivity of various performance quantities with respect to the congestion parameter q (as well as other model parameters). The key insights derived from this sensitivity analysis are: (i) first order effects in general, and the heavy traffic price $\bar{p}$ in particular, are independent of q; (ii) for fixed price, the equilibrium arrival rate and the associated revenue rate are decreasing in q; (iii) the revenue maximizing price is increasing in q; (iv) the equilibrium demand at the revenue maximizing price, the associated congestion cost, and the resulting revenues are decreasing in q.

Numerical results. We now turn to some numerical results that illustrate the accuracy of the proposed approximations. We consider a system with capacity C, potential demand $\Lambda = 2.5 \cdot C$ (connection requests per minute), $\mu = 1$, and a linear demand model with $\alpha = 1 \cdot C$. User sensitivity to congestion is described via the cost $q = \$1$ per unit of lost throughput (per user). Figure 1 depicts three graphs: revenues; price; and, congestion cost as a function of capacity. Each graph has two plots, one describing results for the simulation based optimization, and the other depicting the results obtained via the proposed approximations. Two things, essentially one of the same, stand

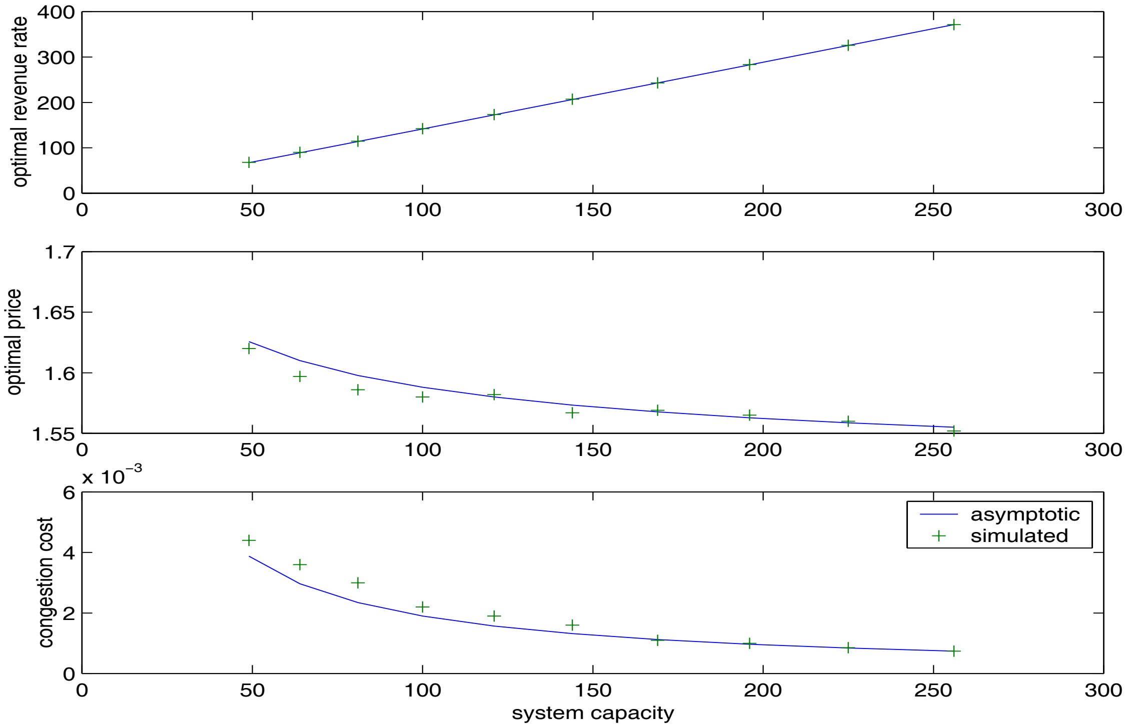

Figure 1: Accuracy of the approximations: (a) optimal revenue rate; (b) the optimal price; and (c) expected congestion cost per user, $q\mathbb{E}D^*$. The " + " plots correspond to the simulation-based optimization results, while the solid line plots correspond to the values derived from the approximate (second order) analysis.

out on close inspection.

First, we observe that the approximation is quite accurate over a range of system sizes. In particular, the optimal price p which was derived via exhaustive simulation-based optimization has the predicted behavior of $p \approx \bar{p} + \pi/\sqrt{C}$, where $\bar{p} = 1.5$ for this problem data. To reiterate this point, the simulation-based optimization did not assume this structure, rather, it emerges as a consequence of the profit maximization objective. The exhaustive search for the optimum is quite tedious given the necessity to seek equilibrium pairs $(\lambda^*(p), q\mathbb{E}D^*(p))$ that satisfy (2). The accuracy of the simple approximation derived from Theorem 2 completely alleviate this computational effort. The second interesting point is not explicitly depicted in the graphs, though it can be inferred in an obvious manner: the "optimized" system operates in "heavy traffic". In fact, it is easy to see that the utilization behaves like $(1 - \rho^*(C)) \approx O(1/\sqrt{C})$.

5 Joint Capacity Sizing and Pricing

So far capacity (C) was taken to be exogenously fixed, or the outcome of some a-priori optimization. We now consider the problem of jointly choosing the system capacity and the price, in order to maximize profits. Inputs to this design problem are the model parameters summarized by the four-tuple (q, Λ, μ, P), and the cost of capacity (appropriately amortized over time), which is assumed to be linear and given by $\$w\mu$ per unit of capacity. That is, we are assuming that there are no economies of scale in the cost structure (in practice, this is typically attributed to increased system complexity).

The monopolistic service provider is faced with the following profit maximization problem:

$$\max_{C,p\geq 0} \quad R(p, C) - wC\mu. \tag{12}$$

Let us denote by C^{rm} and p^{rm} the corresponding maximizers (not necessarily unique). To ensure the service provider can always extract some positive profit by operating such a set of resources, we assume that $P(v > w) = \bar{F}(w) > 0$. Once again, exact solutions are difficult to derive in the absence of a simple characterization of the equilibrium behavior. Our approach targets approximate solutions, and builds on the foundations developed in Sections 3 and 4. The key result was obtained in Theorem 2, where it was shown that for elastic demand ($\varepsilon > 1$) and any choice of capacity C, the revenue maximizing price places the system in the "rationalized" regime, where congestion costs per user are of order $1/\sqrt{C}$, equilibrium demand is $\lambda^* = C\mu - \gamma^*\sqrt{C}\mu$ and the optimal price is of the form $p = \bar{p} + \pi/\sqrt{C}$. The same result holds true when C is a decision variable, provided that the demand elasticity is still valid. Specifically, for any value of capacity C (including the optimal choice C^{rm}), the revenue maximizing price places the system in the "rationalized" regime, where

$$R(p, C) \;\approx\; \left(C\mu - \gamma^*(\pi)\sqrt{C}\mu\right) \times \left(\bar{p}(C) + \frac{\pi}{\sqrt{C}}\right) \;\approx\; C\mu\bar{p}(C) - \sqrt{C}\mu\left(\bar{p}(C)\gamma^*(\pi) - \pi\right) \quad .$$

Substituting the above into the profit maximization problem (12), we arrive at the following approximate formulation:

$$\max_{C\geq 0, \pi\in\mathbb{R}} \quad \underbrace{[C\mu\bar{p}(C) - wC\mu]}_{\text{Capacity sizing}} \;-\; \sqrt{C}\mu\underbrace{\left(\bar{p}(C)\gamma^*(\pi) - \pi\right)}_{\text{Pricing}}.$$

For large systems, the two terms are essentially decoupled suggesting the following heuristic:

(i) *Capacity sizing:* choose the capacity $\hat{C}^{rm}$ that maximizes the nominal profit rate assuming that the system will operate in heavy traffic and neglecting any stochastic effects. This fixes the first order price term $\hat{\bar{p}}^{rm} := \bar{p}(\hat{C}^{rm})$ that places the system in heavy traffic.

(ii) *Pricing:* given the optimal choice for capacity $\hat{C}^{rm}$, choose the second order price component $\hat{\pi}^{rm}$ to minimize the performance degradation due to congestion.

A solution $\hat{C}^{rm}$ to the capacity sizing problem will serve to approximate C^{rm}, the optimal capacity decision, while $\hat{p}^{rm} = \hat{\bar{p}}^{rm} + \hat{\pi}^{rm}/\sqrt{\hat{C}^{rm}}$ will approximate the optimal price p^{rm}.

A more explicit characterization of the approximate solution (given by the decoupling heuristic) can be obtained if we are willing to assume a first order optimality condition for the optimization problem (i). For example, problem (i) is concave in the design variable C for all three demand models mentioned in Section 2. The unique solution satisfies $\hat{C}^{rm}$ in $(0, \tilde{\lambda}/\mu)$, which implies that the increasing costs of capacity force the optimized system to operate in a region where the demand is elastic. The first order condition for an optimizer is

$$\frac{\partial}{\partial C}\left[C\mu\bar{p}(C) - wC\mu\right] \;=\; \mu\bar{p}(C) - \frac{C\mu}{\Lambda}\frac{\mu}{f(\bar{p}(C))} - w\mu \;=\; 0. \tag{13}$$

Recall that $\bar{p}(C) = \bar{F}^{-1}(\kappa)$, where $\kappa = C\mu/\lambda$, thus (13) can be rewritten as $\mu\bar{F}^{-1}(\kappa) - \kappa/f(\bar{F}^{-1}(\kappa)) = w$. Under the aforementioned concavity assumption, this equation has a unique solution denoted by $\hat{\kappa}$, which determines the optimal capacity as a fraction of the potential demand, i.e., $\hat{C}^{rm}\mu = \hat{\kappa}\Lambda$. Since $P(v > w) > 0$, it easy to deduce that $\hat{\kappa} \in (0, 1)$. Given $\hat{\kappa}$, problem (ii) can be used to select the second order price correction term $\hat{\pi}^{rm} = \text{argmin } \{\bar{F}^{-1}(\hat{\kappa})\gamma(\pi) - \pi : \pi \in \mathbb{R}\}$. This is also uniquely determined by $\hat{\kappa}$. In summary, the proposed solution is to set

$$\hat{C}^{rm} = \hat{\kappa}\Lambda/\mu \qquad \text{and} \qquad \hat{p}^{rm} \;=\; \bar{p}(\hat{C}^{rm}) + \frac{\hat{\pi}^{rm}}{\sqrt{\hat{C}^{rm}}} \;=\; \bar{F}^{-1}(\hat{\kappa}) + \frac{\hat{\pi}^{rm}}{\sqrt{\hat{C}^{rm}}} \tag{14}$$

where $\hat{\kappa}$ is determined by (13). The solutions $\hat{C}^{rm}, \hat{p}^{rm}$ scale with Λ according to (14). In the sequel, it will be convenient to recognize this dependence by writing $\hat{C}^{rm}(\Lambda), \hat{p}^{rm}(\Lambda)$. The associated profit rate is $\hat{\mathcal{P}}(\Lambda) = R(\hat{p}^{rm}(\Lambda), \hat{C}^{rm}(\Lambda)) - w\hat{C}^{rm}(\Lambda)\mu$. As the market size realized via the potential demand Λ grows large, the approximate analysis becomes exact and the performance of this heuristic becomes optimal.[10]

Theorem 3 *Let Assumption 1 hold. Then, the revenue rate under the capacity and price pair* $\hat{C}^{rm}(\Lambda), \hat{p}^{rm}(\Lambda)$ *is asymptotically optimal in the sense that*

$$\frac{\hat{\mathcal{P}}(\Lambda)}{\mathcal{P}(\Lambda)} \to 1, \qquad \textit{as } \Lambda \to \infty$$

where $\mathcal{P}(\Lambda) = \max\{R(p(\Lambda), C(\Lambda)) - wC(\Lambda) : C(\Lambda)\mu \geq 0,\; p(\Lambda) \geq 0\}$.

[10] Previous asymptotic results were phrased in terms of C growing large. Given that C is now a decision variable, the natural proxy for system size is the potential demand Λ which reflects the market size.

Given that the profit rate is of the form $[C\mu\bar{p}(C) - wC] - \sqrt{C}\mu(\cdots) + o(\sqrt{C})$, the asymptotic optimality result asserts that the proposed heuristic correctly matches the first order profit term of the optimally designed system. This implies that asymptotically the optimal level of capacity is "close" to $\hat{\kappa}\Lambda$ and that the optimal price is close to $\hat{p}^{rm}$. [11]

Finally, the following numerical example depicted in Figure 2 illustrates the accuracy of this approximate solution. The results are derived for a linear demand model with $\Lambda = 200$ connection requests/min, $\alpha = 50$, $q = 1$, $\mu = 1$ per min and $w = \$1$. For the linear demand model, $\bar{p}(C) = \bar{F}^{-1}\left(\frac{C\mu}{\Lambda}\right) = \frac{\Lambda - C\mu}{\alpha}$. The capacity sizing problem becomes: $\max_{C\geq 0}\{C\mu\frac{\Lambda - C\mu}{\alpha} - wC\mu\}$. The corresponding optimizer is given by $\hat{C}^{rm}\mu = \frac{\Lambda}{2} - \frac{\alpha w}{2} = 75$. Also, $\bar{p}(C^{rm}) = \frac{\Lambda}{2\alpha} + \frac{w}{2} = \$2.5$. Optimizing over the second order term π in the way outlined above gives $\hat{p}^{rm} = \$2.604$. In contrast, the optimal capacity level and price obtained via exhaustive simulation was $C^{rm} = 79$ and $p^{rm} = \$2.529$. The differences are indeed small (and asymptotically negligible).

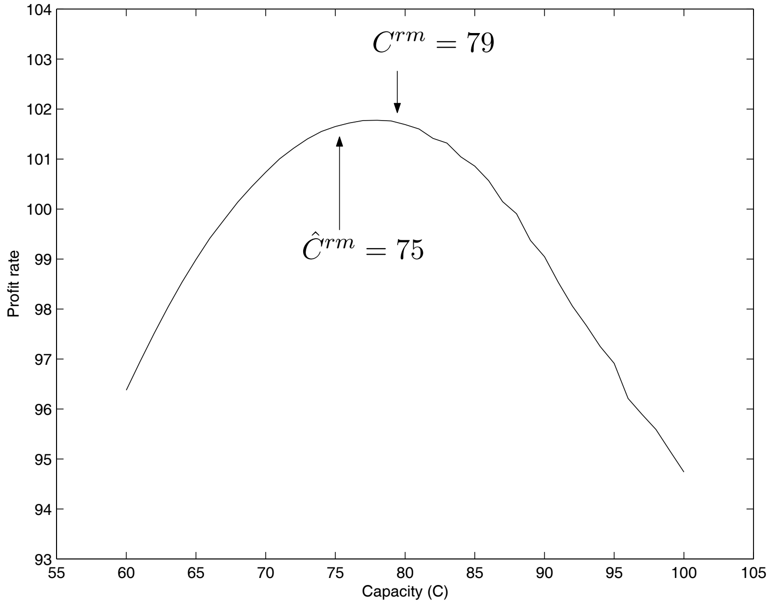

Figure 2: Joint capacity sizing and pricing: $\Lambda = 200$ requests/min, $\alpha = 50$, $q = 1$; $\mu = 1$ per min, $w = \$1$. Decoupled calculation: $\hat{C}^{rm} = 75$ and $\hat{p}^{rm} = \$2.604$. Optimal values obtained via simulation: $C^{rm} = 79$ and $p^{rm} = \$2.529$.

The use of a second order "correction" to control the congestion costs also appears in Borst *et al.* (2000), where this is achieved by adjusting the system's capacity. In contrast to our work, they do not model the choice behavior (demand is fixed) and there is no pricing or equilibrium analysis. To recapitulate, the key insight that emerges is that the capacity sizing and pricing problems decouple

[11] The same approach may still be applicable under more general cost structures, e.g., Theorem 3 can be extended to the case of linear demand and quadratic cost of capacity. The key requirement is that capacity costs do not dominate revenues as capacity grows large.

and both can be easily solved. Capacity is selected to maximize profits to first order, while pricing is selected in order to optimally balance revenues with congestion costs; the optimally sized and priced system operates in heavy traffic.

6 Social Welfare Optimization

We now turn to a brief discussion of large capacity systems that operate under a *social welfare* objective. Since the main results, as well as their derivation, are essentially mirroring the approach taken in the previous section, we will focus here only on a sketch of the main ideas.

Social welfare maximization (social pricing). This formulation assumes that the system is to be operated with the objective of maximizing the total utility. In equilibrium the total expected cost per connection is given by $p + q\mathbb{E}D^*$. Define

$$V(p) := \Lambda \int_{p+q\mathbb{E}D^*}^{\infty} v f(v) dv, \tag{15}$$

to be the *total value* generated per unit time for all subscribers that select to join the system. The *system-wide value* created per unit time is

$$U(p) = \underbrace{V(p) - q\lambda^*\mathbb{E}D^* - p\lambda^*(p)}_{\text{users}} \; + \; \underbrace{p\lambda^*(p)}_{\text{SP}} \; = \; V(p) - q\lambda^*\mathbb{E}D^*, \tag{16}$$

and the service provider's objective is to choose the price p to maximizes the system-wide value.

We proceed using standard arguments to characterize the socially optimal price; see Mendelson and Whang (1990). It is convenient to consider $V(\cdot)$ and $U(\cdot)$ as functions of the rate variable λ, rather than the price p (with slight abuse of notation). The optimal pricing problem will then be phrased in terms of choosing the optimal equilibrium demand rate λ^*, which, in turn, will uniquely define the socially optimal price p. Differentiating U with respect to λ, and evaluating the derivative at the point λ^* (let us denote this as $U'(\lambda^*)$ for simplicity) gives the following first order condition: $U'(\lambda^*) = V'(\lambda^*) - q\mathbb{E}D^* - \lambda^* q \left.\dfrac{\partial \mathbb{E}D^*(\lambda)}{\partial \lambda}\right|_{\lambda^*} = 0$. From $\lambda^* = \Lambda \bar{F}\left(p + q\mathbb{E}D^*\right)$, it follows that the utility of the marginal user that joins the system when the equilibrium demand rate is λ^* is $V'(\lambda^*) = p + q\mathbb{E}D^* = \bar{F}^{-1}(\lambda^*/\Lambda)$. Substituting this in the expression above gives the socially optimal price, p^{soc}

$$p^{soc} = \lambda^* q \frac{\partial \mathbb{E}D^*}{\partial \lambda^*}. \tag{17}$$

That is, it is socially optimal to charge each user the externality (or congestion) cost that he imposes on the system. Exact evaluation of the socially optimal price is again quite complicated due to the equilibrium formulation, and thus relies on exhaustive simulation or numerical approximation of the steady-state equilibrium probabilities. As we show next, the corresponding asymptotic characterization is much simpler to work with, and the task of social optimization is computationally trivial.

An argument similar to that of Section 4 shows that maximizing social welfare will also "drive" the system to operate in the "rationalized" regime. Counterparts to Proposition 3 and Theorem 2 can be derived under the same modelling assumptions with one change: here we do not require demand elasticity. [12] Using the scaling relations derived in Theorem 1 we can derive the appropriate social optimization objective for the asymptotic system, and proceed with the second order analysis to approximate the optimal price. In the following derivation we will use the fact that $\mathbb{E}D^* = d^*/\sqrt{C} + o(1/\sqrt{C})$, where $d^* := d(\gamma^*)$; see Section 3. Starting from (16) we can write

$$\begin{aligned} U(\lambda^*) &\approx \Lambda \int_{\bar{p}+\frac{1}{\sqrt{C}}(\pi+qd^*)}^{\infty} v f(v) dv - (C\mu - \gamma\sqrt{C}\mu) q \frac{d^*}{\sqrt{C}} \\ &\approx \Lambda \int_{\bar{p}}^{\infty} v f(v) dv - \frac{1}{\sqrt{C}} \Lambda \bar{p} f_\tau(\bar{p})(\pi + qd^*) - q\mu\sqrt{C} d^* \\ &= V(C\mu) - \sqrt{C}\mu \left(\bar{p} \frac{f(\bar{p})}{\bar{F}(\bar{p})}(\pi + qd^*) + qd^* \right). \end{aligned} \tag{18}$$

Since $\gamma = \frac{f(\bar{p})}{\bar{F}(\bar{p})}(\pi + qd^*)$, asymptotically the social optimization problem reduces to choosing the second order price term π to $\min_{\pi \in \mathbb{R}} \ \{\bar{p}\gamma^*(\pi) + qd^*(\gamma^*(\pi))\}$, making the dependence on γ^* and π explicit. Recall also that $d(\gamma) = \nu(\gamma)/\gamma$. This is readily solvable given the characterization of the equilibrium $\gamma^*(\pi)$ in (8). Taking derivatives with respect to γ, the first order optimality condition is given by $\partial/\partial\gamma^* \, (\bar{p}\gamma^* + qd^*) = 0$, which implies that

$$\bar{p} = -q \left. \frac{\partial d(\gamma)}{\partial \gamma} \right|_{\gamma^*}. \tag{19}$$

Note the similarity between (19) and (11) derived under the revenue maximization objective.

Discussion. Expression (19) is closely related to (17), and has a natural interpretation. Recall that the optimal pricing rule is of the form $p = \bar{p} + \frac{\pi}{\sqrt{C}}$, thus for very large capacity systems each subscriber pays $\bar{p}$. The right-hand-side of (19) is the *externality cost*, where the negative sign accounts for the fact that as γ^* increases the arrival rate into the system decreases. Hence, the socially optimal equilibrium corresponds to the operating point γ^{soc} where the externality cost $-q\frac{\partial d^*}{\partial \gamma^*}$ equals the first order price $\bar{p}$ paid by the users. Parenthetically, it follows from (11) that

[12] In the absence of congestion costs the social value function $U(\cdot)$ becomes $U(p) = \Lambda \int_p^\infty v f(v) dv$, which is decreasing in p. That is, as we lower price and $\lambda(p)$ increases, the system-wide value increases. This characteristic of the social value function is essentially the equivalent of the elasticity assumption and will drive the economically optimized system to heavy traffic.

under revenue maximization the service provider charges a fixed premium over the externality cost. Moreover, using (19) and the explicit expression for $d^* = d(\gamma^*)$ given in (7) we can evaluate the term $\frac{\partial d^*}{\partial \gamma^*}$ as a function of γ^*, and solve for the socially optimal equilibrium denoted by γ^{soc}. This characterization together with the definition of the equilibrium in (8) specify the socially optimal price, viz, $p^{soc} = \bar{p} + \pi^{soc}/\sqrt{C}$, where $\pi^{soc} = \left(\bar{F}(\bar{p})/f(\bar{p})\right)\gamma^{soc} - qd(\gamma^{soc})$.

Joint capacity sizing and social pricing. This problem is treated in an identical manner to the revenue maximization counterpart. The objective is now to solve the problem

$$\max_{C\geq 0, p\geq 0} \quad \left(\Lambda \int_{p+q\mathbb{E}D^*}^{\infty} vf(v)dv - \lambda^* q\mathbb{E}D^*\right) - wC\mu,$$

which leads to the following asymptotic formulation:

$$\max_{C\geq 0, \pi\in\mathbb{R}} \quad \underbrace{\left[\Lambda \int_{\bar{p}(C)}^{\infty} vf(v)dv - wC\mu\right]}_{\text{Capacity sizing}} - \underbrace{\sqrt{C}\mu\left(\bar{p}(C)\gamma^*(\pi) + q\mathbb{E}\mathcal{D}^*\right)}_{\text{Pricing}}.$$

For large systems, this problem therefore decouples into two parts: (i) choose the capacity to maximize the social surplus; and (ii) choose the price π to minimize performance degradation due to congestion. Problem (i) is a simple maximization of a concave function. (This follows from the the general assumptions on the choice model, and the form of the heavy-traffic price $\bar{p} = \bar{F}^{-1}(\kappa)$.) Exploiting the decoupling of the capacity sizing and pricing decisions, we can compare the optimal capacity levels and prices for the approximating revenue and social optimization problems. Let $\hat{C}^{rm}$, $\hat{C}^{soc}$ and $\hat{p}^{rm} = \bar{p}^{rm} + \frac{\pi^{rm}}{\sqrt{C}}$, $\hat{p}^{soc} = \bar{p}^{soc} + \frac{\pi^{soc}}{\sqrt{C}}$ be the optimal capacities and prices for the two approximating problems, essentially optimizing the respective objectives up to and including second order terms. Then, simple analysis shows that

$$\hat{C}^{soc} > \hat{C}^{rm} \quad \text{and} \quad \bar{p}^{soc} < \bar{p}^{rm} \quad . \tag{20}$$

Moreover, one can show that asymptotically as the market size grows large, this implies that the actual optimal capacity and price decisions will also be ordered in the same way, i.e., $C^{soc} > C^{rm}$ and $p^{soc} < p^{rm}$. That is, the socially optimal solution will have higher capacity and charge less than the revenue maximizing one, which is consistent with similar comparisons made in the context of various other systems dating back to the seminal paper by Naor (1969).

7 Concluding Remarks

Motivated by the proliferation of communication and information services, this paper introduces and analyzes a model for systems with large capacity, and resources that can be shared among users when the system is congested but cannot be pooled when the system is underutilized. Such systems exhibit *statistical economies of scale* in the sense that they become more "efficient" as their capacity increases. A tractable asymptotic analysis leads to several structural insights.

(i) Operating regime: The system should operate close to "heavy traffic," where nominally all resources are fully utilized. This is optimal in the context of revenue maximization (under the assumption that demand is elastic), and under a social optimization objective.

(ii) Pricing: The optimal price admits a simple decomposition; the service provider computes a price to bring the system to "heavy traffic," and subsequently applies a second order correction term that depends on the size of the system, to optimally balance congestion effects.

(iii) Joint capacity and pricing decisions: Given the natural scaling relationships intrinsic to such systems, the capacity sizing and pricing problems decouple. Capacity is chosen to maximize profits (or total system value) assuming that the system is fully utilized and neglecting stochastic variability. The price is then adjusted to optimally balance congestion costs.

Several interesting directions of future research arise. One natural extension concerns a system that supports *differentiated services*. In the information and communication service context, in particular, for systems that share resources, the natural first step is to consider a menu with two service grades: "guaranteed" and "best effort." The former refers to users that are *guaranteed* a constant rate of service irrespective of the state of the system, while the latter refers to users that share the remaining capacity (and thus are prone to service degradation). To this end, the Halfin-Whitt many-server asymptotic regime supports certain diffusion approximations that can be used to facilitate the study differentiated services [some preliminary results along these lines are derived in Maglaras and Zeevi (2002)].

While this paper has focused on steady-state analysis, the diffusion approximations that were pioneered by Halfin and Whitt (1981) could be used to approximate transient behavior in such systems, adding another important layer to the current static analysis. Finally, an even more challenging problem concerns dynamic pricing mechanism that extend the static fixed-price setting considered herein. In particular, the results in the current paper strongly suggest that this would revolve around second-order analysis as well.

A Proofs

For notational clarity we will denote the capacity of the system as $C_n = n$. Since most of our results concern the asymptotic regime where C (thus, n) grows arbitrarily large, various variables and quantities will be appropriately indexed by a subscript n. The proofs of Propositions 1 and 3 are omitted and are available in a technical report version of the paper.

Proof of Theorem 1: The convergence of the sequence of equilibrium traffic intensities follows from Proposition 1 in Halfin and Whitt (1981). First note that the Markov chain associated with the system under investigation in this paper is identical to that associated with an $M/M/n$ system. Halfin and Whitt (1981) considered a sequence of $M/M/n$ queues and showed that $\lim_{n\to\infty} \mathbb{P}(N_n \geq n) = \nu \in (0,1)$ if and only if $\lim_{n\to\infty} \sqrt{n}(1-\rho_n) = \gamma > 0$, where $\nu(\gamma) = \phi(\gamma)/(\gamma\Phi(\gamma)+\phi(\gamma))$. Applying their result to our sequence of systems (with capacities C_n) operating in equilibrium, we get that $\mathbb{P}(\text{congestion}) \to \nu \in (0,1)$ implies that $\sqrt{n}(1-\rho_n^*) \to \gamma^* > 0$. By Proposition 1 in Halfin and Whitt (1981), it also follows that the two conditions are equivalent. The next step is to establish the limit for the expected congestion cost (part (iii)); later on, this will be combined with the structure of the customer choice model to derive the asymptotic decomposition of the pricing rule (part (ii)).

The starting point is the relation for the expected excess delay given in (9), which we repeat here for completeness $\mathbb{E}D_n^* = \rho\mathbb{P}(N_n^* \geq n)/(n(1-\rho_n^*))$. From here, using Halfin and Whitt (1981), it follows that if $\sqrt{n}(1-\rho_n^*) \to \gamma^*$, then $\sqrt{n}\mathbb{E}D_n^* \to d(\gamma^*) := \nu(\gamma^*)/\gamma^*$.

In the sequel we will also use the notation $d(\gamma)$ when we want to make explicit its dependence on γ. The right hand side in the above expression is $d(\gamma^*)$ appearing in part (iii) of the statement of the main result. Finally, we will use (i) and (iii) to establish the desired price decomposition. For n sufficiently large, (iii) implies that $\mathbb{E}D_n^* = d/\sqrt{n} + o(1/\sqrt{n})$. Using this expansion, Taylor's theorem and the smoothness of the choice distribution in the definition of λ_n^*, we get that

$$\begin{aligned}
\lambda_n^* &= \Lambda_n P\left(v \geq p_n + q\mathbb{E}D_n^*\right) \\
&= \Lambda_n P\left(v \geq p_n\right) - \Lambda_n f(p_n) q \frac{d}{\sqrt{n}} + o(\sqrt{n}) \\
&= n\mu - \gamma^*\sqrt{n}\mu + o(\sqrt{n}),
\end{aligned}$$

where the last equality follows from (i), i.e., $\sqrt{n}(1-\rho_n^*) \to \gamma^*$. Now, by assumption $\Lambda_n = n\mu\kappa^{-1}$, thus the last equality implies that $\Lambda_n P\left(v \geq p_n\right) = n\mu + \delta\sqrt{n}\mu + o(\sqrt{n})$, for some appropriate choice of $\delta \in \mathbb{R}$. This, in turn, implies that the price p_n must be of the form $p_n = \bar{p} + \pi/\sqrt{n} + o(1/\sqrt{n})$, where $\bar{p}$ is selected such that $P(v \geq \bar{p}) = \kappa$, that is, $\bar{p} = \bar{F}^{-1}(\kappa)$. This establishes part (ii) of the Theorem. Finally, using this structural form of p_n, Taylor's theorem, and the smoothness of the choice model distribution we can express λ_n^* in the form

$$\lambda_n^* = \Lambda_n P(v \geq \bar{p}) - \Lambda_n \frac{f(\bar{p})}{\sqrt{n}} \left(\pi + qd + o(\sqrt{n})\right) + o(\sqrt{n}). \tag{21}$$

Using $\lambda_n = n\mu\kappa^{-1}$ we have in the limit as $n \to \infty$ that $\gamma^* = (\pi + qd(\gamma^*))\, f(\bar{p})/\bar{F}(\bar{p})$. ■

Proof of Proposition 2: Consider any converging subsequence $\{n_j\}$ such that $\sqrt{n_j}(1-\rho_{n_j}^*) \to \gamma_j > 0$. As noted in the proof of Theorem 1, we have that $\sqrt{n_j}\mathbb{E}D_{n_j}^* \to d(\gamma_j) = \nu(\gamma_j)/\gamma_j$. Next, we take a Taylor expansion of the expression for $\lambda_{n_j}^*$

$$\begin{aligned}\lambda_{n_j}^* &= \Lambda_{n_j} P\left(v \geq p_n + q\mathbb{E}D_{n_j}^*\right)\\ &= \Lambda_{n_j} P(v \geq \bar{p}) - \frac{1}{\sqrt{n_j}} \Lambda_{n_j} f(\bar{p})\left(\pi + qd(\gamma_j)\right) + o(\sqrt{n_j}).\end{aligned}$$

Rewrite Λ_{n_j} as $n_j\mu/\bar{F}(\bar{p})$ to get that $\lambda_{n_j}^* = n\mu - \sqrt{n_j}\mu\frac{f(\bar{p})}{\bar{F}(\bar{p})}(\pi + qd(\gamma_j)) + o(\sqrt{n_j})$, which implies that $\gamma_j = (f(\bar{p})/\bar{F}(\bar{p}))\,(\pi + qd(\gamma_j))$. Solving for $d(\gamma_j)$ we get that

$$d(\gamma_j) = \frac{1}{q}\left(\frac{\bar{F}(\bar{p})}{f(\bar{p})}\gamma_j - \pi\right). \tag{22}$$

The expression for $d(\gamma_j)$ and (22) imply that the equilibrium γ_j must satisfy (8). Now, observe that (8) can be rewritten as $h(\gamma) := q^{-1}\bar{F}(\bar{p})/f(\bar{p})\gamma - \gamma^{-1}\phi(\gamma)(\gamma\Phi(\gamma)+\phi(\gamma)) = q^{-1}\pi$. By inspection $h(\cdot)$ is a continuous increasing function on the positive half-line, and $h(0) = -\infty$. [This follows since $\phi(\gamma)$ is a decreasing function on the positive half line, while $\Phi(\gamma)$ is increasing.] Thus, $h(\gamma) = q^{-1}\pi$ has a unique solution $\gamma^*(\pi)$ for all $\pi \in \mathbb{R}$. Consequently (8) has a unique solution $\gamma^*(\pi)$ which implies that all converging subsequences have the same limit $\gamma^*(\pi)$. ■

Proof of Theorem 2: The main idea is to establish that the "rationalized regime" leads to profit maximization, and then appeal to Theorem 1 in Section 3 to conclude that the price structure must be of the form asserted in the current theorem. The only added task is to establish that the second order price correction π is indeed the solution to the given optimization problem, stated in the theorem. The proof is divided into three steps.

Step 1. Proposition 3 asserts that $\rho_n^{rm} \to 1$ as $n \to \infty$. We will now determine the rate of this convergence, which in turn will imply that $p^{rm} \to \bar{p}$. To this end, consider the price sequence $\{p_n^{rm}\}$ that is revenue maximizing, i.e., $p_n^{rm} \in \operatorname{argmax}_p\{R(p,n)\}$ for each $n \geq 1$. Here $R(p,n)$ extends the definition in (3) denoting the revenue rate for a system with capacity $C_n = n$, and making this dependence explicit. In what follows, all system quantities are considered in equilibrium, and we omit the '*' superscript for notational clarity. Let us denote the resulting expected excess delay by $d_n := \mathbb{E}D_n^*$. Recall from (9) we have that $d_n = \rho_n\mathbb{P}(N_n^* \geq n)/(n(1-\rho_n))$. Suppose that $\rho_n^{rm} \to 1$ so that $\liminf_{n\to\infty} n(1-\rho_n^{rm}) \leq M$, for some finite positive constant $M > q\mu/\bar{p}$. Then, it must be that $\sqrt{n}(1-\rho_n^{rm})=o(1)$, where $a_n=o(1)$ if $a_n\to 0$ as $n\to\infty$. By Proposition 1 of Halfin and Whitt (1981) we then have that $\mathbb{P}(N_n^*\geq n)\to 1$ and thus $\limsup_{n\to\infty} d_n\geq M^{-1}$. Now,

$$\frac{\lambda_n}{n\mu}=\frac{\Lambda_n\bar{F}(p_n^{rm}+qd_n)}{n\mu}\to 1,$$

thus, it follows that $p_n^{rm}+qd_n\to\bar{p}$. But since $d_n\geq M^{-1}$ infinitely often, it follows that $\liminf_{n\to\infty} p_n^{rm}\leq\bar{p}-q/M$ and consequently

$$\liminf_{n\to\infty}\frac{R(p_n^{rm},n)}{R(\bar{p},n)}\leq 1-\frac{q}{M\bar{p}}\ ,$$

where the right hand side is strictly less than 1 by choice the M. Therefore, p_n^{rm} cannot be the revenue maximizing price, in contradiction. We conclude that for any positive constant M, $\liminf_{n\to\infty} n(1-\rho_n^{rm})\geq M$, and therefore since M is arbitrary we must have $n(1-\rho_n^{rm})\to\infty$. Thus, $d_n=o(1)$ and this yields that $p_n^{rm}\to\bar{p}$.

Step 2. Here we establish that $p_n^{rm}\to\bar{p}$ in combination with the demand elasticity assumption imply that $\sqrt{n}(1-\rho_n^{rm})\to\gamma$. First, since $\lambda_n=n\mu(1-\gamma_n)$ we have

$$\lambda_n=n\mu-\frac{n\mu f(\bar{p})}{\bar{F}(\bar{p})}(\pi_n+d_n)+O\left(n^2(\pi_n+qd_n)^2\right), \tag{23}$$

where $\gamma_n:=1-\rho_n^{rm}$, and $\pi_n=p_n^{rm}-\bar{p}$. [The second equality above follows from the equilibrium condition that "ties together" d_n,π_n and γ_n, see (22)]. Thus, we can express the revenue rate as follows

$$R(p_n^{rm},n)=\lambda_n p_n^{rm}=n\mu\bar{p}-\underbrace{n\mu\left(\bar{p}\gamma_n-\pi_n\right)}_{\psi(\gamma_n)}+O(n\pi_n\gamma_n)$$

where the last term on the right hand side is of lower order. Taking a closer look at the second order term above, and using (23) we have that

$$\begin{aligned}\psi(\gamma_n)&=n\mu\left(\bar{p}\gamma_n-\pi_n\right)\\&=n\mu\left(\gamma_n\bar{p}-\frac{\bar{F}(\bar{p})}{f(\bar{p})}\gamma_n+qd_n\right)\\&=\sqrt{n}\mu\left[\sqrt{n}\gamma_n\left(\bar{p}-\frac{\bar{F}(\bar{p})}{f(\bar{p})}\right)+qd_n\sqrt{n}\right]\end{aligned}\tag{24}$$

Now, if $\sqrt{n}\gamma_n=o(1)$, then from (9) and Proposition 1 of Halfin and Whitt (1981) which asserts that $\mathbb{P}(N_n^*\geq n)\to 1$, we have that $\sqrt{n}d_n\to\infty$. Consequently, $\psi(\gamma_n)/\sqrt{n}\to\infty$. On the other hand, if $\sqrt{n}\gamma_n\to\infty$ then by Proposition 1 of Halfin and Whitt (1981) and using (9), it follows that $\limsup_{n\to\infty}\sqrt{n}d_n<\infty$. In addition, the demand elasticity assumption implies

$$\varepsilon\ =\ -\frac{\partial\Lambda\bar{F}(p)}{\partial p}\frac{p}{\Lambda\bar{F}(p)}\ =\ f(p)\frac{p}{\bar{F}(p)}\ >\ 1$$

thus, it follows that $\bar{F}(p) < pf(p)$. Consequently, $\psi(\gamma_n)/\sqrt{n} \to \infty$. Now, if $\sqrt{n}\gamma_n \to \gamma > 0$, then $\psi(\gamma_n)/\sqrt{n} \to c > 0$, which minimizes the rate of growth of the lost revenues, due to second order effects. Thus, $1/\sqrt{n}$ is the economically optimal rate of convergence to heavy-traffic.

Step 3. To conclude the proof we again appeal to Proposition 1 of Halfin and Whitt (1981) which asserts that if $\sqrt{n}(1-\rho_n^{rm}) \to 1$, then $\mathbb{P}(N_n^* \geq n) \to \nu \in (0,1)$. Consequently, the "rationalized regime" is the economically optimal regime, viz, supporting the best possible rate of growth of revenue (up to second order effects). Moreover, by Theorem 1 it follows that the price sequence corresponding to this regime is of the form $p_n^{rm} = \bar{p} + \frac{\pi}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$. Thus, $\pi_n = \pi/\sqrt{n}$. Finally, the optimal choice of π is obtained by minimizing the effects of the second order term $\psi(\gamma_n)$. Spelling this out, using (24), we have $\pi^{rm} = \operatorname{argmin}_{\pi \in \mathbb{R}}\{\gamma^*(\pi)\bar{p} - \pi\}$, where here we have made explicit the fact that γ depends on π (through (8)). ■

Proof of Theorem 3: Consider the profit maximization problem max $\{R(p(\Lambda), C(\Lambda)) - wC(\Lambda)\mu : C(\Lambda), p(\Lambda) \geq 0\}$, and define $\kappa(\Lambda) := C(\Lambda)\mu/\Lambda$. We will prove that $\kappa(\Lambda) \to \hat{\kappa} \in (0,1)$ as $\Lambda \to \infty$, where $\hat{\kappa}$ is defined via $\hat{C} = \hat{\kappa}\Lambda/\mu$, the unique root of (13). It then follows that the decoupled heuristic of (14) is asymptotically optimal.

To shorten notation, the superscript "rm" will be dropped in the sequel. First, recall that $\hat{\kappa} \in (0,1)$ and that $\hat{C}(\Lambda)\mu/\Lambda = \hat{\kappa}$ for all Λ. Also, note that $\bar{p}(\hat{C}(\Lambda)) = \bar{F}_\tau^{-1}(\hat{\kappa}) := \bar{p}(\hat{\kappa})$. It is easy to show that under the proposed heuristic $\frac{1}{\Lambda}[R(\hat{p}(\Lambda), \hat{C}(\Lambda)) - w\hat{C}(\Lambda)\mu] \to \hat{\kappa}\bar{p}(\hat{\kappa}) - w\hat{\kappa} > 0$.

Let's assume that $\kappa(\Lambda) \to \kappa \in [0,1]$. If $\kappa = 0$, then $\frac{1}{\Lambda}[R(p(\Lambda), C(\Lambda)) - wC(\Lambda)\mu] \to 0$, which contradicts the optimality of $(p(\Lambda), C(\Lambda))$. Hence $\kappa > 0$. From Section 4 we know that systems with large capacity $C(\Lambda)$ under the revenue maximizing price operate in heavy traffic and their profit rate can be expressed in the following form: $C(\Lambda)\mu\bar{p}(C(\Lambda)) - wC(\Lambda\mu - \sqrt{C(\Lambda)}\mu(\cdots) + o(\sqrt{C(\Lambda)})$. It follows that

$$\frac{1}{\Lambda}[R(p(\Lambda), C(\Lambda)) - wC(\Lambda)\mu] \to \kappa\bar{p}(\kappa) - w\kappa,$$

where $\bar{p}(\kappa) := \lim_{\Lambda\to\infty} \bar{p}(C(\Lambda))$. Since $\hat{\kappa}$ is assumed to be the unique optimizer of the capacity sizing problem, i.e., it is the uniquely defined via (13), it follows that $\kappa\bar{p}(\kappa) - w\kappa \leq \hat{\kappa}\bar{p}(\hat{\kappa}) - w\hat{\kappa}$, with equality achieved only when $\kappa = \hat{\kappa}$. In order to establish that $\kappa = \hat{\kappa}$ we need to ensure that this asymptotic profit rate is achieved by the optimal policy. This is clearly true since our proposed heuristic achieves that rate, and the optimal policy can only do better. It follows that $\kappa = \hat{\kappa}$. Using the asymptotic expansions for the profit rates under the optimal policy and under the heuristic of (14) we immediately establish the asymptotic optimality result. ■

References

[1] M. Armony and C. Maglaras. On customer contact centers with the call-back option. *Preprint*, 2001.

[2] S. Borst, A. Mandelbaum, A. and M. Reiman. Dimensioning of Large Call Centers. *preprint*, 2000.

[3] C. Courcoubetis, F. Kelly and R. Weber. Measurement-based usage charges in communications networks. *Oper. Res.*, **48**:535-548, 2000.

[4] A. Das and R. Srikant. Diffusion approximations for a single node accessed by congestion controlled sources, *IEEE Trans. on Aut. Control*, **45:**1783-1799, 2000.

[5] G. Gallego and G. van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Mngt. Sci.*, **40**:999-1020, 1994.

[6] O. Garnett, A. Mandelbaum and M. Reiman. Designing a call center with impatient customers. *M&SOM*, **4**:208-227, 2002.

[7] R. Gibbens and F. Kelly. Resource pricing and the evolution of congestion control. *Auromatica*, **35**:1969–1985, 1999.

[8] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Op. Res.*, **29**:567-588, 1981.

[9] J. M. Harrison. A broader view of Brownian networks. To appear *Ann. Appl. Prob.*, 2002.

[10] S. Lanning, D. Mitra, Q. Wang and M. Wright. Optimal planning for optical transport networks. *Phil. Trans. Royal Soc. London A*, **1773**:2183–2196, 2000.

[11] C. Maglaras and A. Zeevi. Diffusion approximations for a Markovian service system with "guaranteed" and "best effort" service levels. Working paper, Columbia Univeristy, 2002.

[12] J. McGill and G. van Ryzin. Revenue management: research overview and prospects. *Trans. Sci.*, **33**:233-256, 1999.

[13] H. Mendelson. Pricing computer services:queueing effects. *Com. ACM*, **28**:312-321, 1985.

[14] H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Oper. Res.* **38**:870–883, 1990.

[15] P. Naor. On the regulation of queue size by levying tolls. *Econometrica*, **37**:15-24, 1969.

[16] I. Paschalidis and J. Tsitsilkilis. Congestion-dependent pricing of network services. *IEEE/ACM Trans. on Networking*, **8**:171-184, 2000.

[17] S. Stidham. Optimal control of admission to a queueing system. *IEEE Trans. on Aut. Control*, **30**:705-713, 1985.

[18] J. Van Mieghem. Price and service discrimination in queueing systems: incentive compatability of G$c\mu$ scheduling. *Mngt. Sci.*, **46**:1249-1267, 2000.

[19] W. Whitt. Understanding the efficiency of multi-server service systems. *Mngt. Sci.*, **38**:708-723, 1992.

[20] W. Whitt. How multiserver queues scale with growing congestion-dependent demand. To appear in *Op. Res.*, 2003.

Addendum: Additional Proofs

Proof of Proposition 1: Fix a capacity $C_n = n$, and price $p > 0$, such that $P(v > p) > 0$ (otherwise there is nothing to prove). Fix $\xi > 0$ and, with some abuse of notation, let the expected excess delay for an arrival rate $\lambda(p + q\xi)$ be denoted by $\mathbb{E}D(\xi)$. The equilibrium demand rate and expected excess delay are now defined via the solution ξ^* of the set of equations

$$\lambda(p + q\xi^*) \;=\; \Lambda P(v > p + q\xi^*) \qquad \text{and} \qquad \xi^* \;=\; \mathbb{E}D(\xi^*),$$

if such a solution exists. To this end, let $h(\xi) = \xi - \mathbb{E}[D(\xi)]$. Now, λ is a decreasing function when considered w.r.t. the variable ξ, and ξ (the expected excess delay per-user) is, in turn, increasing in λ. This suggests that there exists a fixed point ξ^* that solves the above equations. To make this rigorous, note that $h'(\xi) = 1 - \frac{\partial}{\partial \xi}\mathbb{E}D(\xi) > 0$, since $\mathbb{E}D(\xi)$ is decreasing in ξ. Differentiability of $\mathbb{E}D(\xi)$ follows from two observations. First, $\mathbb{E}D^*$, considered here as a function of λ^*, is continuously differentiable in $[0, n\mu)$. This follows from (9) and the expressions for the steady-state probability of congestion, $\mathbb{P}(N \geq n)$, for an M/M/n queue [see, e.g, Halfin and Whitt (1981)]. Second, $\lambda(\cdot)$ is continuously differentiable due to the smoothness of the choice model. Thus, by the chain rule, $\mathbb{E}D(\xi)$ is continuous and differentiable. Since $h(0) < 0$ and $h(\infty) > 0$, it follows that $h(\xi) = 0$ has a unique solution ξ^*, and the associated variables $\lambda^* := \lambda(p + q\xi^*)$ and $\mathbb{E}D(\xi^*) = \xi^*$ characterize the unique equilibrium of the finite capacity system. Finally, the traffic intensity is then given by $\rho^* := \lambda^*/(n\mu)$ which is strictly less than 1. ■

Proof of Proposition 3. First, by Proposition 1 we have that $\rho_n^* = \lambda^*/(n\mu) < 1$ for all $n \geq 1$. Thus,

$$\limsup_{n\to\infty} \rho_n^* \leq 1 \tag{25}$$

Consider a price sequence $\{p_n^{rm}\}$ that is profit maximizing, i.e., $p_n^{rm} \in \operatorname{argmax}_p R(p, n)$ for each n, where $R(p, n)$ extends the definition in (3) denoting the revenue rate for a system with capacity $C_n = n$. Recall that $\Lambda_n := n\mu/\kappa$. Now, suppose that $\liminf_{n\to\infty} \rho_n^{rm} \leq 1 - \delta$ for some $\delta \in (0, 1)$, where ρ_n^{rm} is the utilization in a system with capacity n under the price p_n^{rm}, in equilibrium. Then,

$$\begin{aligned}
\rho_n^{rm} &= \frac{\Lambda_n}{n\mu} P(v > p_n^{rm} + q\mathbb{E}D_n) \\
&= \kappa^{-1} P(v > p_n^{rm} + q\mathbb{E}D_n) \\
&= \kappa^{-1} P(v > p_n^{rm}) - \kappa^{-1} f(p_n^{rm}) q\mathbb{E}D_n + O\left([\mathbb{E}D_n]^2\right) .
\end{aligned}$$

Using (9) we have that $\mathbb{E}D_n \leq (1-\delta)/n\delta$ for infinitely many n, thus $\liminf_{n\to\infty} \mathbb{E}D_n = 0$. From here it follows that $\liminf_{n\to\infty} \kappa^{-1}\bar{F}(p^{rm}) \leq 1-\delta$. Consequently, we have that $P(v > p_n^{rm}) \leq \kappa(1-\delta/2)$, say, for infinitely many n. Consider now the price sequence $\tilde{p}_n = \bar{p}$, where $\bar{p} = \bar{F}^{-1}(\kappa)$, i.e., the price that places the system in heavy-traffic. The assumptions on the choice model together with the above imply that for some $\delta' > 0$ we have $p_n^{rm} \geq \tilde{p}_n(1+\delta')$ for infinitely many n. Now, due to demand elasticity we have $x\bar{F}(x) \uparrow$ as $x \downarrow$, and thus, using the assumed smoothness of the choice model, we have that for some $\delta'' > 0$,

$$\limsup_{n\to\infty} \frac{R(\tilde{p}_n, n)}{R(p_n^{rm}, n)} \geq 1 + \delta'',$$

in contradiction to the claimed optimality of $\{p_n^{rm}\}$. Thus, it must be that $\liminf_{n\to\infty} \rho_n^{rm} \geq 1 - \delta$, and since δ was arbitrary, this implies that for the profit maximizing price sequence $\liminf_{n\to\infty} \rho_n^{rm} \geq 1$, which together with (25) establishes that $\rho_n^{rm} \to 1$ under the profit maximization objective. This concludes the proof. ■