Pricing and Performance Analysis for a System with Differentiated Services and Customer Choice

Constantinos Maglaras
Graduate School of Business
Columbia University
NY, NY 10027
c.maglaras@columbia.edu

Assaf Zeevi
Graduate School of Business
Columbia University
NY, NY 10027
assaf@gsb.columbia.edu

Abstract

We consider a model of a service system with finite and shared processing capacity and two service classes. Users arrive at the system and select either a high-priority service level where the service requests are processed at a fixed rate, or a low-priority service level where service rate is subject to degradation when the system is congested. A fixed price-per-connection is charged for each service level, and the mean delay in each class class is announced to the users upon arrival. The users, in turn, select the appropriate class of service based on their perceived "cost," comprised of price and delay-related cost. We demonstrate that the optimal operational mode of this system is in "heavy-traffic" if the demand is elastic, and determine the asymptotically optimal price per service grade. In particular, the magnitude of price-premium for high-priority service is seen to be "small." Finally, a somewhat surprising feature of the system is that the fraction of users that select each service-level is determined by a "second-order" analysis that hinges on underlying diffusion limits.

1 Introduction

The recent proliferation of web-based services has triggered service providers to explore ways by which to address processing requirements of diverse applications and concomitantly segment the market of potential users. That is, service providers are attempting to offer multiple grades of service so that users are differentiated according to their quality-of-service (QoS) requirements, and willingness-to-pay for desired QoS.

This paper proposes a simple and tractable model for service systems that offer *differentiated* and *substitutable services*, i.e., where the user can select between service levels each time s/he accesses the system. In particular, we consider a service provider (SP) that operates a system with fixed capacity and offers two classes of service. A high-priority service (HP) grants users a fixed processing rate, but users may have to wait to access the service if the system is congested. A low-priority service (LP) where users access service immediately, however the processing rate is subject to degradation. In particular, if there

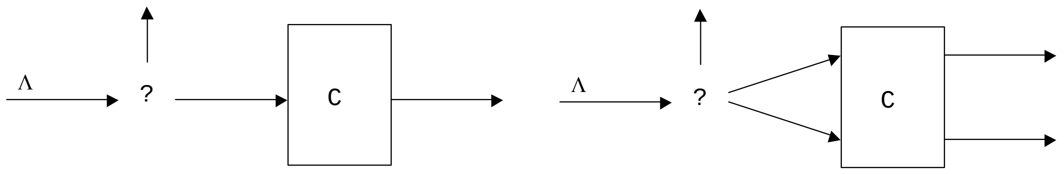

Figure 1: Schematic model representation: one user population accesses a system with either one (left model) or two -differentiated- service grades (right model); the paper focuses on the latter.

is enough processing capacity to allocate to the low-priority class, these users will each receive a nominal rate allocation; otherwise, processing capacity allocated to this class of service is split among these users in an egalitarian manner, and consequently users experience congestion in the form of a degraded service rate. This somewhat stylized configuration in which users select the service grade based on congestion, which in turn serves as a feedback signal in regulating system usage and user choice, is reminiscent of the the Paris Metro pricing scheme; cf. [7].

As explained in more detail in the subsequent section, users who connect to the system select a service grade based on their price and congestion sensitivity. In this sense, the basic services delivered to users are *substitutable*, and the users merely select the *quality grade* according to which the service will be delivered. A schematic representation of the system under investigation is given in Figure 1. The service provider is then faced with the objective of maximizing profit by selecting the optimal fixed prices for each service level. The main thrust of this paper is to perform an economic analysis that includes the determination of the optimal prices, and to assess their implications on the performance of the system (e.g., congestion levels, utilization, fraction of users joining each service level etc.). The key economic assumption that we impose is that the demand function, determined by the user choice model, is *elastic*. This supposition is in line with empirical findings for information/communication services; see [8].

Exact analysis of the system described above is difficult to carry out even under simplifying Markovian assumptions, due to the intrinsic congestion “feedback” mechanism that governs the choice behavior of users. (The system steady-state behavior is then governed by the *equilibrium* induced by this feedback.) What we pursue in this paper is an analysis that hinges on asymptotic approximations which are applicable in systems with large processing capacities that handle high volumes of connection requests. The main underlying machinery that facilitates analysis in this regime is given by diffusion approximations that capture the system dynamics and support closed form derivations, including a computation of the system equilibrium.

The main contributions of this paper are summarized in the following insights.

1. The optimal pricing schedule results in high system utilization. In particular, as

the system capacity (C) and potential demand both grow large, the utilization of processing resources (ρ) increases at a rate $\rho = 1 - \gamma/\sqrt{C} + o(1/\sqrt{C})$ as $C \to \infty$. (See Proposition 2.)

2. The optimal prices for each grade of service are asymptotically equal to a common value, that being the price that induces full utilization of system resources. (See Proposition 3.)

3. The system equilibrates in the so-called Halfin-Whitt heavy-traffic regime, where congestion in the low-priority class is of order $O(1/\sqrt{C})$, and congestion in the high-priority class is negligible (exponentially small). Finally, the equilibrium operating point is given by the solution to a fixed point equation, and can be computed efficiently. (See Theorem 1.)

4. The price-premium for priority service (i.e., HP-grade) is "second order," that is, $p_1 - p_2 = O(1/\sqrt{C})$. Moreover, the fraction of incoming users joining each service grade is determined by a "second order" analysis, in particular, it depends on the price premium for priority. (See Theorem 1.)

5. If the optimal operating point "shuts-off" the low-priority class, the resulting system operates as single class service with delays that are of order $O(1/\sqrt{C})$. (See Corollary 1.)

We note that while there may be instances (given by specific choice model parameters) that lead to optimal economic performance by essentially "shutting-off" the LP-class, roughly speaking, the operation of the two-class system will typically lead to revenues that are strictly higher than those extracted using a single class system.

The model that we posit with capacity constraints and shared resources is closely related to the one studied by Das and Srikant [2] and Maglaras and Zeevi [4] in a single-class setting. An analysis of a system with two grades of service without choice and where the services are not substitutable was carried out recently by Maglaras and Zeevi [5]. The "equilibrium" formulation, where steady-state congestion signals serve to regulate the system and user behavior follows the work of Mendelson and Whang [6]. The main ideas in the analysis are inspired by the work of Halfin and Whitt [3] that concerns queueing systems with "many" servers.

The remainder of the paper is structured as follows. Section 2 describes the underlying model and Section 3 pursues the economic optimization objective using the asymptotic approximations alluded to earlier in this section. We conclude in Section 4 where we discuss the main insights that arise from this analysis, in particular, the effect of customer choice on the performance of the system.

2 The System Model and Problem Formulation

System model and service grades. Our system model has a finite processing capacity, denoted by C, which can be considered as processing rate that is allocated to the users. Two service grades are offered to incoming connection requests: a high-priority (HP) and low-priority (LP) grade. Let $N_1(t)$ and $N_2(t)$ denote the number of users in the system at time $t \geq 0$ in the HP-class and LP-class, respectively. (This subscript convention will be used to tag various other quantities to associate them with the two service grades.) The HP-grade processes users' requests at a fixed unit rate when $N_1(t) < C$, and users that arrive when $N_1(t) \geq C$ wait in a queue until a unit of processing capacity becomes available. We use $D_1(t)$ to denote the queueing time experienced by a user arriving at time t. The LP-grade processes users' requests at a fixed unit rate when $N_2(t) < C - N_1(t)$, i.e., when residual capacity not allocated to the HP-class affords an allocation of unit rate to each user in the LP-class. When $N_2(t) \geq C - N_1(t)$, users in the LP-class receive a degraded processing rate which is equal to $[C - N_1(t)]/N_2(t)$, i.e., when resources are congested, users in the LP-class share processing resources allocated to that grade of service in an egalitarian manner. To facilitate analysis, we take as a proxy for the actual excess delay (due to rate degradation) faced by a user that selects LP-grade service at time $t \geq 0$

$$D_2(t) = \begin{cases} 0 & N_1(t) + N_2(t) \leq C \\ m\left(\frac{N_2(t)}{C - N_1(t)} - 1\right) & N_1(t) + N_2(t) > C, \end{cases}$$

where $m = 1/\mu$ is the mean processing time for user requests (explained below). In large capacity systems, $D_2(t)$ turns out to be an asymptotically correct approximation to the actual excess delay due to a pathwise version of Little's law. Finally, the service provider levies a fixed per-user price-per-connection charge p_1, p_2 for each service grade respectively.

User choice behavior. Connection requests arrive according to a Poisson process with rate Λ. This rate can be thought of as the *market potential* for the offered service; this interpretation will be useful in the analysis that follows in section 3. Users have random processing requirements that are i.i.d. exponentially distributed random variables with mean $m = 1/\mu$, and independent of the Poisson arrival process. Each user has a valuation v for the service and a delay sensitivity parameter q, he observes the mean congestion levels in the system, namely, $\mathbb{E}D_i(t)$ for $i = 1, 2$ (this information is announced by the system), and then evaluates his utility according to

$$u_i(t) = v - p_i - q(m + \mathbb{E}D_i(t)) \qquad i = 1, 2.$$

The type of each user is determined by the pair of parameters v, q that are assumed to be random, and follow a joint distribution F with continuous density f, and i.i.d. across users, independent of the arrival and service time processes. If $u_i \geq \max\{0, u_j\}$, $i, j = 1, 2$ and $i \neq j$, then the incoming user will select service grade i, and thus the rate at which users join each service is given by

$$\lambda_i(p_i,t) = \Lambda\mathbb{P}(u_i(t) \geq 0,\ u_i(t) > u_j(t)), \quad i,j = 1,2,\ \ i \neq j$$

for each service grade respectively. Our analysis restricts attention to the class of *demand functions* λ, as determined by the user choice model, that are elastic. For our model this is best described by assuming, hypothetically, that both service classes are priced at a common value $\$p$, i.e., $p_1 = p_2 = p$, and that both result in the same total delay $d = m + d_i$ for $i = 1, 2$. In this case, the total arrival into the system would be $\lambda(p;d) = \Lambda\mathbb{P}(v - qd \geq p)$. Keeping d fixed, the demand function $\lambda(p;d)$ is said to be *elastic* at the price p if

$$\varepsilon(p) := -\frac{\partial\lambda(p;d)}{\partial p}\frac{p}{\lambda(p;d)},$$

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

Assumption 1 The demand function $\lambda(p;d)$ is elastic in $\{p : \lambda(p;d) \in [0, C\mu],\ d \geq m\}$.

Intuitively, a demand function is elastic if a decrease in price (and increase in demand) result in an increase in the revenue rate $p\lambda(p;d)$. (The elasticity of the demand function is determined by the characteristics of the users, namely, the distribution of valuations for service and delay sensitivity (v,q).)

Equilibrium formulation. 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 $\lambda = (\lambda_1, \lambda_2)$ and corresponding congestion cost $d = (\mathbb{E}D_1, \mathbb{E}D_2)$ such that both are time-independent and jointly satisfy the demand relationship

$$\lambda_i(p) := \Lambda\mathbb{P}(u_i \geq 0,\ u_i > u_j), \quad i,j = 1,2,\ \ i \neq j\ . \tag{1}$$

where $u_i = v - p_i - q(m - qd_i)$, and $d_i := \mathbb{E}D_i(\infty)$ denotes the steady-state mean excess delay. 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 $N = (N_1(t), N_2(t) : t \geq 0)$, such the expected excess delay w.r.t. to this distribution, $\mathbb{E}D_i(\infty)$, $i = 1, 2$, induces a time homogenous external arrival rate $\lambda = (\lambda_1, \lambda_2)$ through (1), and λ, in turn, is consistent with the aforementioned steady-state distribution. The next result establishes the existence and uniqueness of this equilibrium regime (Proofs are omitted from the paper due to space limitations.)

Proposition 1 *For each capacity $C > 0$, and price vector $p \geq 0$, there exists a unique steady-state equilibrium.*

Finally, the fraction of entering traffic that selects the low priority service option is given by

$$\kappa := \frac{\lambda_2(p)}{\lambda_1(p) + \lambda_2(p)}.$$

Economic optimization objective. We assume that the service provider operates in a market with imperfect competition, where she can influence the demand rate by changing the price menu, and that she has perfect knowledge of the user type (v, q) distribution F and the mean service requirement per user m. Given a system with capacity C, the service provider's objective is to select the price vector p to maximize the equilibrium revenue rate given by

$$R(p) = p \cdot \lambda(p), \tag{2}$$

where $a{\cdot}b$ denotes the inner product between two vectors. Implicit in (2) is the dependence of revenues on the congestion in the system, which in turn is affected by the price schedule p.

We note that the system model imposes implicitly a non-idling assumption on the service disciplines, i.e., that service provider cannot intentionally idle resources when there are users (mainly LP users) in the system that require service. Also, the service provider cannot introduce any extraneous delay to the LP class, e.g., by introducing a delay node for the LP users after their processing task is completed but before they are made aware of that fact. (This would apply in services that are performed remotely and where the user cannot observe the work in progress but only the end result.)

3 Economic Analysis and System Performance

This section has two main objectives. The first is to study the revenue maximization problem formulated in section 2. The second objective is that of performance analysis, where we seek to develop a set of tools to analyze the behavior of a system with a given capacity C, where the price vector p need not necessarily be optimal. In the interest of space, our approach will to focus here on the former objective, and in passing we will develop an approach to analyze the latter as well.

The main idea that underlies our analysis is an asymptotic framework that focuses on a regime where capacity, C, and market potential, Λ, both grow large, while remaining proportional to each other. In addition, the user characteristics, summarized in terms of their type distribution F and their mean processing requirement, m, are held fixed. That is, our analysis looks at large systems that serve a large user market with characteristic described in the previous section. To explicitly denote the connection between C and Λ we will write the latter as $\Lambda = C\bar{\Lambda}$, where $\bar{\Lambda}$ is a normalized market size per unit of offered capacity. To illustrate this asymptotic regime, consider a given model with, say, 100 units of capacity and market potential equal to 400. Our analysis then proceeds to examine the asymptotic behavior of a sequence of systems with capacity C and potential demand $\Lambda = 4C$, as C grew large. The behavior of the original system would then be extracted by interpreting the limiting results for $C = 100$, thus effectively "evaluating" the limiting system at the capacity level that corresponds to the original system of interest.

The first step is to note that the rates at which users join each service class can be expressed as

$$\begin{aligned}\lambda_1(p) &= \Lambda\mathbb{P}\left(q \geq \frac{p_1 - p_2}{d_2 - d_1},\ v - q(m + d_1) \geq p_1\right),\\ \lambda_2(p) &= \Lambda\mathbb{P}\left(q < \frac{p_1 - p_2}{d_2 - d_1},\ v - q(m + d_2) \geq p_2\right),\end{aligned}$$

where the events whose probability is evaluated above correspond to the conditions: $u_1 \geq u_2$ and $u_2 > u_1$, respectively, and $u_i \geq 0$ for $i = 1, 2$. It is also natural to posit that $p_1 \geq p_2$ and $d_2 \geq d_1$, given the nature of each service grade (and this can indeed be verified).

In the sequel, let p_i, ρ and $d_i := \mathbb{E}D_i$, for $i = 1, 2$, denote the revenue maximizing price vector, the equilibrium resource utilization rate, and the corresponding excess delays. By attaching a superscript 'c' to various quantities, e.g., the abovementioned variable, we denote their dependence on the system capacity C, which will grow large. The main results of this section study these economically optimal quantities under the demand elasticity assumption (Assumption 1).

Our first result indicates that when capacity and demand both grow large proportionally to each other, it is optimal to operate the system in a regime where resources are almost fully utilized and where the HP class suffers negligible queueing-related congestion.

Proposition 2 *$\rho^c \to 1$ and $\mathbb{E}D_1^c \to 0$ as $C \to \infty$.*

Assuming optimistically that $\mathbb{E}D_2^c \to 0$ as $C \to \infty$, we can define the price $\bar{p}$ that would induce full resource utilization as follows

$$\Lambda\mathbb{P}(v - qm \geq \bar{p}) = C\mu\ . \tag{3}$$

Note that the left hand side of (3) is equal to $\lambda_1(p) + \lambda_2(p)$, contingent on $\mathbb{E}D_1 = \mathbb{E}D_2 = 0$. The next result establishes that the optimal prices in each service grade converge asymptotically to $\bar{p}$, as C grows large.

Proposition 3 *$p_i^c \to \bar{p}$, for $i = 1, 2$, as $C \to \infty$.*

The proposition implicitly implies that $\mathbb{E}D_2^c \to 0$ as $C \to \infty$.

The main result of this section describes in precise detail the manner in which an economically optimized system approaches a regime where resources are almost fully utilized, i.e., a heavy-traffic operating point. Moreover, the analysis provides a closed form characterization of the asymptotic system behavior, viz., equilibrium, congestion etc. In the sequel the notation $a^c \approx b^c$ is used to denote equality to within $o(1/\sqrt{C})$ terms, i.e., $a^c = b^c + o(1/\sqrt{C})$.

Theorem 1 *As $C \to \infty$,*

i.) *Resource utilization:* $\rho^c \approx 1 - \dfrac{\gamma}{\sqrt{C}}$

[ii.) *Congestion:* $\limsup_{C\to\infty} e^{B(\kappa)C}\mathbb{E}D_1^c < \infty$ *and* $\mathbb{E}D_2^c \approx \dfrac{d(\gamma)}{\kappa\sqrt{C}}$*, for* $B(\kappa) > 0$*, and* $\kappa = \lim_{C\to\infty} \lambda_2^c/(\lambda_1^c + \lambda_2^c)$.

iii.) *Optimal pricing strategy:* $p_i^c \approx \bar{p} + \dfrac{\pi_i}{\sqrt{C}}$*, where*

$$(\pi_1, \pi_2) \in \operatorname{argmin}\left\{\bar{p}\gamma - \bar{\kappa}\pi_1 - \kappa\pi_2 \ : \ \pi_2 \leq \pi_1\right\}.$$

iv.) *The system equilibrium: is the unique solution to*

$$\gamma = a_1\pi_1 + a_2\pi_2 + a_3 d(\gamma)/\kappa,$$

where $d(\gamma) = \phi(\gamma)[\gamma(\gamma\Phi(\gamma) + \phi(\gamma)]^{-1}$*, and* a_1, a_2, a_3 *are explicit function related to derivatives of the user choice distribution* F*, and identified explicitly below.*

From Theorem 1 we immediately obtain

Corollary 1 *Under the conditions of Theorem 1, if* $\pi_1 = \pi_2$*, then* $\kappa = 1$ *and* $\bar{\kappa} = 0$*, and the system reduces to a single-class one where all users receive HP service. The delay in that class, namely,* $\mathbb{E}D_1^c$*, is such that* $\mathbb{E}D_1^c \approx d(\gamma)/\sqrt{C}$.

We note that the proofs of Theorem 1 and Corollary 1 hinge on diffusion limits that are derive for the state processes which track the number of users in the system in each service class. Some structural insights that are gleaned from this theorem are discussed in the next section.

We conclude our discussion by specifying the constants a_1, a_2, a_3 that appear in the equilibrium equation above, and justifying formally the optimization problem that determines the price corrections π_1, π_2. Assuming that $p_i^C = \bar{p} + \pi_i/\sqrt{C}$, $\mathbb{E}D_1^c \approx 0$ and that $\mathbb{E}D_2^c \approx \delta_2/\sqrt{C}$ for $\delta_2 = d(\gamma)/\kappa$, we have that

$$\lambda_1^c \ = \ \Lambda\mathbb{P}\left(q \geq \frac{\pi_1 - \pi_2}{\delta_2}, \ v - qm \geq \bar{p} + \pi_1/\sqrt{C}\right) \ \approx \ \bar{\kappa}C\mu - \sqrt{C}\mu a_1\pi_1,$$

where $\bar{\kappa} = 1 - \kappa$ and

$$\kappa = \frac{\mathbb{P}(q \leq \frac{\pi_1 - \pi_2}{\delta_2}, \ v - qm \geq \bar{p})}{\mathbb{P}(v - qm \geq \bar{p})}, \quad \text{and} \quad a_1 = -\frac{\partial[\mathbb{P}(q \leq \frac{\pi_1-\pi_2}{\delta_2}, \ v - qm \geq \bar{p})]}{\partial\bar{p}} \cdot \frac{1}{\mathbb{P}(v - qm \geq \bar{p})}.$$

Similarly, for the LP class we have that

$$\lambda_2^c \ = \ \Lambda\mathbb{P}\left(q < \frac{\pi_1 - \pi_2}{\delta_2}, \ v - q(m + \delta_2/\sqrt{C} \geq \bar{p} + \pi_1/\sqrt{C}\right) \ \approx \ \bar{\kappa}C\mu - \sqrt{C}\mu(a_2\pi_2 + a_3\delta_2),$$

where

$$a_2 = -\frac{\partial[\mathbb{P}(q < \frac{\pi_1-\pi_2}{\delta_2},\ v - qm \geq \bar{p})]}{\partial \bar{p}} \cdot \frac{1}{\mathbb{P}(v - qm \geq \bar{p})}$$

and

$$a_3 = -\frac{\partial[\mathbb{P}(q < \frac{\pi_1-\pi_2}{\delta_2},\ v - qm \geq \bar{p})]}{\partial m} \cdot \frac{1}{\mathbb{P}(v - qm \geq \bar{p})}.$$

Summing $\lambda_1^c + \lambda_2^c$ and dividing by $C\mu$ we get the expression for γ that characterizes the equilibrium. To establish that the equilibrium is unique we need to relate the equilibrium to an implicit equation in terms of κ, which is then shown to have a unique solution. Finally, it is easy to verify that the total revenue rate has the form

$$\lambda^c \cdot p^c = \bar{p}C\mu - \sqrt{C}\mu\left(\bar{p}\gamma - \bar{\kappa}\pi_1 - \kappa\pi_2\right) + o(\sqrt{C}) \quad \text{as } C \to \infty.$$

This asymptotic form of the revenue rate which provides a scale decomposition into first order and second order terms, provides formal (non-rigorous) justification of the optimization problem used to determine the optimal price corrections π_1 and π_2 in Theorem 1(iii.).

4 Discussion

The economic optimality of the heavy-traffic operating regime, in particular, the Halfin-Whitt regime, is in some sense not surprising given the analysis of a single class system performed in [4]. In that paper, demand elasticity is the key assumption in deriving this optimality result. Specifically, the service provider can extract higher revenues by lowering prices, which in turn leads to higher demand and increased resource utilization. The results of Halfin and Whitt [3] indicate that in a single class $M/M/N$ queue, where N increases and the traffic intensity approaches 1, utilization can be close to 100% while delays are still “small.” This is precisely the regime where the system equilibrates in our formulation.

Theorem 1 can be used for purposes of performance analysis: specifically, given any price vector p, once can re-write this in the form $p = \bar{p} + \pi_i/\sqrt{C}$ for appropriate π_i, $i = 1, 2$, where $\bar{p}$ is the price that induces full resource utilization and C is the system capacity. The theorem can then be used to to characterize the equilibrium behavior under this pricing scheme, which can be computed using a simple numerical procedure. This, in turn, can be used to analyze the generated revenues. This approach is in line with the view that the system with finite capacity C can be “embedded” within the asymptotic derived in Theorem 1.

We note that the “first order” behavior of the system, in particular, the demand rate for each class of service, is determined by a second order analysis (which dictates the choice of π’s and results in second order congestion effects). That is, κ which captures the fraction of incoming requests that select each service grade, is determined by the system equilibrium, which in turn hinges on second order parameters (e.g., γ, which measures the second order arrival rates into the system). This stands in stark contrast to most of the heavy-traffic literature where the first order behavior of the system can be derived based on a "deterministic," fluid limit, analysis. In [5] this is shown to be the case in a model where two user populations access two service classes without choice. Thus, one fundamental feature of the model with choice is this somewhat unusual dependence of "first order" system performance on "second order" parameters.

Finally, we note that by "injecting" idleness in a judicious manner in the LP-class, users, perceiving the extra delay as part of the nominal service rate degradation due to the congestion effects, might be elicited to pay more for the HP service option. (This point has been raised recently by Afeche [1] in the context of a single-server queueing node.) The use of this mechanism as a means by which to increase revenues is explored in detail in the full version of this paper, where it is shown that one can optimize the level of "injected" idleness through a simple deterministic optimization problem that hinges on the asymptotic analysis pursued in this paper.

References

[1] P. Afeche. Optimal strategic idleness in queueing systems: Maximizing revenues by doing nothing. 2002. Working paper, Kellogg School of Management.

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

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

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

[5] C. Maglaras and A. Zeevi. Pricing and design of differentiated services: Approximate analysis and structural insights. 2003. Working paper, Columbia Univeristy.

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

[7] A. Odlyzko. Paris metro pricing for the internet. *In Proc. ACM Conf. on Elec. Comm.*, pages 140–147, 1999.

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