A call center with both contract and non-contract customers was giving priority to the contract customers only in off-peak hours, precisely when having priority was least important. In this paper we investigate whether this is rational behavior on the part of the call center and what the implications are for customers. In particular, we show that under contracts on the percentile of delay, which are commonly used in the call-center industry, this is rational behavior, at least under the asymptotic regime considered in the paper. We then suggest other contracts that do not result in this type of undesirable behavior from a contract customer’s perspective. We compare the performance of the different contracts in terms of mean, variance, and outer percentiles of delay for both customer types using both numerical and asymptotic heavy-traffic analyses. We argue that including terms reflecting the second moment of delay in a contract would be beneficial to contract customers and, in a sense, fairer.
1 Introduction

Call (or contact) centers are in many ways the public face of a firm. Indeed, the impact of poor service from a call center can resonate throughout an organization. However, increasingly firms are outsourcing their call center needs to other organizations, notably call centers located in distant locales from the home of the firm (see, e.g., Global Call Center Outsourcing Summit, 2004). Such outsourced call centers often serve more than one customer. Further, many such call centers serve customers of firms who have contracted for service with the call center as well as non-contract customers who may pay for services on an ad-hoc basis. In order to ensure quality from outsourced call centers, firms sign service level agreements which dictate terms by which a call center will be measured; within these agreements there are terms which dictate acceptable customer delays.

One such service level contract dictates a so-called 80/20 rule, meaning in this case 80% of all customers must be served within 20 seconds (see, e.g., Jackson 2002). Of course, the particular numbers may vary, but the form is quite common (see, e.g., White Paper On Service Level Agreements, 2000). Failure to meet terms of these contracts can result in a service level credit that would either result in a rebate to the contract customer or credit for future service (see, e.g., OutsourcingBestPractices.com, 2001). We will refer to this type of contract, where customers have been guaranteed that delay does not exceed some fixed time more than some given percentile of time, as a delay percentile contract.

Call center contracts are not limited to firms outsourcing their customer/consumer call center needs. Such contracts are also used by firms seeking support for their own staff. For example, firms contract with call centers for computer software support for their employees. The important point is that call centers are used to address a wide variety of clients and, in many cases, some but not all of the clients have service level agreements that must be met. Further, call centers have the ability, through both interactive menus and caller-ID recognition software, to determine the level of support guaranteed to callers.

The motivating example for this paper is a call center that gave priority to contract customers operating under a delay percentile contract only in off-peak hours, the time when having priority was least important. At other times, the firm gave priority to non-contract customers, increasing the likelihood of their return while serving contract customers only to the extent that their contract was fulfilled. In this paper we investigate whether such a policy is rational and the implications of such contracts for call center management. We do not investigate the outsourcing decision (see, e.g., Akşin et al.(2005) and references therein).

We consider a stylized model of a call center (the firm) that serves both contract and non-contract customers. In our model contract customers have a service level agreement that the firm must adhere to. Initially we consider a case where the call center firm seeks to minimize the waiting time of non-contract
customers while adhering to the service level contract constraint. By doing so, the firm hopes to improve the service experience for these customers and retain them for future business. (See e.g. Cross, Liedtka, and Weis, 2005.) This was the objective of the motivating example for this paper. This would also be the objective of a call center supporting outsourced information technology that generated additional revenue from non-contract, pay-as-you-go, customers. If one considers the call center as a firm to be a rational economic agent, such a model is reasonable. As discussed in Gans et al. (2003), abandonment probabilities are often linearly related to expected delay; thus, in some sense, minimizing expected delay can be seen as a proxy for minimizing abandonment probabilities for non-contract customers. Further, we show that this model is equivalent to one where the firm is assessed linear penalties for failure to meet the contract as is sometimes done in practice (Taylor, 2005).

To analyze the model, we approximate the system by its heavy-traffic, steady-state scaled limit (a concept we formalize later) and derive the optimal policy for the firm for the limiting problem. We then present a regular traffic policy designed to result in the heavy-traffic policy in the limit. In particular, we show that the policy, while meeting the contracted percentile for delay, actually has some properties that are highly undesirable from the contracting customer’s perspective and may not have been expected at the time the contract was signed. We show that the delay percentile contract, as used in practice, can result in long delays and high coefficients of variation for contract customers relative to what they might have achieved without a contract and under First-Come-First-Serve service. In order to address these deficiencies, we propose several alternate contracts.

Practitioners advocate the use of contracts which increase penalties for increased delays and repeated failures to meet agreed upon service levels (Boyle and Tochner, 2004). To model such contracts we consider a contract that assesses a convex penalty on the firm for delaying the contract customers and a contract that assesses a convex penalty on contract customers delayed beyond the contractual limit while ensuring the contract is maintained on average. Finally, we consider a percentile delay contract with a quadratic penalty if the waiting time for contract customers exceeds that of the non-contract customers. Such a contract, while not observed in practice, might result from a contract customer learning of their increased delays vis-à-vis non-contract customers. We analyze and discuss the performance of these contracts.

The literature on call center is vast and growing. Gans et al. (2003) provides an excellent tutorial and literature review on the topic. Further, Mandelbaum (2004) provides a recent bibliography on the topic, which contains over 400 references. Given these two comprehensive references, we only provide a limited literature review on call center research here; we restrict attention to call center papers we feel are directly relevant to our work. In particular, we focus is on the so-called “V designs” described in Gans et al. (2003).

The most related work to ours is probably an as yet unpublished note by Koole (2005), which we became
aware of after our study. He considers a single class of customers and argues that delay percentile contracts cause a rational provider to ignore customers that have exceeded their delay, while a convex metric would induce the provider to serve customers with the longest delay first. Performance analysis and implementation issues are discussed. Note that he is looking at scheduling within a class, whereas our study examines the trade-off between two classes of customers, namely contract and non-contract customers.

Three related papers that consider a similar idea of contract and non-contract customers are Bhulai and Koole (2003), Gans and Zhou (2003), and Saltzman and Mehrotra (2001). In all of these papers the contract customers have a constraint on some function of delay, while the non-contract customers are “best-effort.” In Bhulai and Koole (2003) and Gans and Zhou (2003) the objective is to maximize throughput of non-contract customers, who are always available, while meeting the delay constraint for contract customers. On the other hand, Saltzman and Mehrotra (2001) do not perform policy optimization. Instead they design a simulation model to evaluate staffing levels in a situation where contract customers are given the guarantee that they will wait less than one-minute or the call is free. Further literature on more tangentially related two-type customer papers is given in Gans and Zhou (2003).

Because this paper uses heavy-traffic theory to develop the contracts, both for the original delay percentile contract and the alternative suggested contracts, it is related to the large stream of literature that consider asymptotic optimality in heavy-traffic. We note that we refer to this literature as it motivates the limiting control problem that is the focus of our paper. (As such our emphasis is not on demonstrating the existence of the asymptotic policy.) For example, it is related to Van Mieghem (1995) and Mandelbaum and Stolyar (2004) who consider scheduling of multiple classes of customers with convex delay costs on homogeneous and heterogenous servers, respectively. However, neither of these papers have constraints on service. (See Mandelbaum and Stolyar 2004 for further related literature when there are no constraints.)

Closer to the work in this paper, Plambeck et al. (2001) consider the scheduling of a single server queueing system with multiple delay classes where there are delay constraints. They also consider admission control. A highly related paper is Maglaras and Van Mieghem (2004), who consider a fluid model rather than the heavy-traffic diffusion model considered in Plambeck et al. (2001). The reader is referred to Maglaras and Van Mieghem (2004) for further literature on systems with delay constraints.

All the previously cited heavy-traffic literature consider systems scaled by “conventional” heavy-traffic scalings. Of perhaps more relevance to call centers is the so-called Halfin-Whitt or QED regime first proposed by Halfin and Whitt (1981). Authors considering asymptotic optimality of scheduling policies under the Halfin-Whitt regime include: Atar (2004, 2005), Atar et al. (2004a, 2004b), Armony (2004), Armony and Maglaras (2004a, 2004b), Harrison and Zeevi (2004), and Gurvich (2004).

Of the papers in the previous paragraph, only Armony and Maglaras (2004a, 2004b) and Gurvich (2004)
consider delay constraints. Like Maglaras and Van Mieghem (2004) and Plambeck et al. (2001), Armony and Maglaras (2004a, 2004b) assume that such constraints must be “asymptotically compliant”, which implies that the delay constraints are met 100% of the time in the heavy-traffic limit. Instead, we only require that the delay requirements are met some given percentage of the time in steady-state. The only other work we are aware of that considers steady-state constraints is Gurvich (2004) who has constraints on the percentage of customers that experience any delay.

While not considering constraints on delay, another related work is that of Bassamboo, Harrison and Zeevi (2004) that considers the problem of minimizing staffing costs and customer abandonment costs for multiple customer and server types. Of relevance to our work is their proposed asymptotic parameter regime in which the limiting system equilibrates instantly, allowing the control problem they solve to be tractable. Another related paper is that of Wallace and Whitt (2005) that also considers the staffing and control problem for multiple customer types where each customer type has different service level agreements. They study a simulation-based heuristic for solving the problem under the assumption of a fixed priority.

We feel that the most important contributions of this paper relate to its practical implications for call center contracts. However, another contribution of this paper is our notion of asymptotic optimality, which will be seen to be significantly different to previous papers in the literature. We show that depending on the contract type, alternate threshold-type policies are used in heavy-traffic. However, these policies differ considerably from the policies found in the papers above. This is because both the objective and constraint differ from that traditionally used in the heavy-traffic literature. As outlined above, delays are usually guaranteed to be less than some fixed value in the scaled system. In order for a delay percentile contract (as commonly used in industry) to make sense under heavy-traffic scalings, we must express the constraints in terms of the expectation of the delay. This requires the development of a new type of asymptotic optimality that deals with steady-state expectations.

This paper is organized as follows. Section 2 presents the basic model and assumptions. It outlines the heavy-traffic theory necessary for the paper and gives the optimization model for the firm. Section 3 presents the main results of the paper. The optimization model is solved and a “regular traffic” policy presented that is argued to converge to the optimal policy in heavy-traffic. This policy is shown to be undesirable from the contract customers’ perspective and alternate policy forms are discussed. Finally, Section 4 concludes the paper with discussion on the results and possible avenues for future research.
2 Model and Assumptions

We consider a call center with two classes of customer, namely contract (type 1) and non-contract (type 2) customers. Customers within each class are homogeneous and are assumed to arrive to queue $i$ according to a general renewal process with independent identically distributed interarrival times with mean $1/\lambda_i$ and squared coefficient of variation $c_{si}^2$, $i = 1, 2$. Let $\Lambda = \lambda_1 + \lambda_2$ be the total arrival rate and $c_a^2 = (\lambda_1 c_{s1}^2 + \lambda_2 c_{s2}^2)/\Lambda$ be the averaged squared coefficient of variation over both arrival processes. Also let $p_i = \lambda_i / \Lambda$ be the proportion of arrivals to class $i$, $i = 1, 2$. The arrival rates for both classes are assumed to be constant; thus, we do not examine the problem of gaining and/or losing customers. Also, we do not consider abandonment from the queue. Both extensions are discussed briefly in Section 4.

There are $N$ servers that work in parallel. Both types of customer are assumed to have the same service time distribution (we discuss relaxation of this assumption later). Further, service times have the same distribution across all servers with mean $1/\mu$ and squared coefficient of variation $c_s^2$. Let $\rho_i = \lambda_i / (N\mu)$ be the traffic intensity of class $i$ customers and let $\rho = \rho_1 + \rho_2$ be the system traffic intensity; we assume that $\rho < 1$. All interarrival times and service times are assumed to be mutually independent and independent of the state of the system.

The delay percentile contract that we consider is specified by two parameters, $D$ and $\gamma$. The contract requires that $\gamma \times 100\%$ of all contract (type 1) customers wait no more than $D$ time units. A policy defines at each time point which server is working on which queue. A feasible policy must satisfy the contract when the time-average over all departed customers is considered. We restrict attention to feasible policies that are First-Come-First-Serve (FCFS) within customer class. Such policies are often referred to as “Head-of-the-Line” service; we return to discussion of this assumption in Section 4. Finally, we only consider policies that are stable, in that queue length and delay (see below) have proper limiting distributions in a time-average sense (see, e.g., p. 237, Wolff 1989) yielding steady-state random variables.

Returning to notational definitions, let $X_i(t)$ be the number of customers of type $i$ in the system at time $t$, $i = 1, 2, t \geq 0$. Further, let $X(t) = X_1(t) + X_2(t)$ be the total number of customers in the system. Note that $X(t)$ is completely determined by the arrival and service input processes and the policy chosen. Similarly, let $Q_i(t)$ be the number of customers of type $i$ in the queue (not in service) and $Q(t) = Q_1(t) + Q_2(t)$. Note that $Q(t) = (X(t) - N)^+$, where $x^+$ is the positive part of some variable $x$. Let $Q_1$, $Q_2$, and $Q$, be the corresponding steady-state random variables for $Q_1(t)$, $Q_2(t)$, and $Q(t)$, respectively, under some given stable policy.

Let $W_i(t)$ be the “virtual” delay in queue for a type $i$ customer at time $t \geq 0$, $i = 1, 2$. That is, $W_i(t)$ is the delay until service a type $i$ customer would receive if such a customer arrived at time $t$. Let $W_i$ have
the steady-state distribution of \( W_i(t) \) under some given stable policy. When arrivals are Poisson then the PASTA property (see, e.g., p. 293, Wolff 1989) implies that \( W_i \) is also steady-state delay. Under heavy-traffic it will turn out that \( W_i \) still represents steady-state delay, even for non-Poisson arrival processes.

In this paper, rather than analyzing the actual system in terms of customer delay or queue length, we will concentrate on analysis of the system under heavy-traffic. As such we rely on an asymptotic analysis which removes the detailed system dynamics of the queue that may be of interest to call center managers. While analysis of policies for the system in “regular” traffic (either the transient or steady-state solutions) is difficult, considerable simplifications occur under heavy-traffic. We can thus obtain tractable results which will then provide insights that will be applied to the original regular traffic system.

At this point we make no other assumptions on policies. However, in general, heavy-traffic limits require that policies be work conserving or non-idling (see, e.g., p. 69, Whitt 2002). Such a condition means that at any given time, along the same input sample paths, the amount of work in the system (which is the time to complete all jobs if no more arrivals occur) is constant across this policy class. Policies are also frequently required to be non-anticipatory, in that decisions must depend only on information that has been revealed at the current time and not use future, yet to be revealed, values of the processes.

The rest of this section is organized as follows. Section 2.1 provides definitions for the heavy-traffic limits for the system under consideration in this paper and discusses important concepts in such limits, namely “state-space collapse” and “steady-state convergence”. We give examples of these definitions in Section 2.2. In Section 2.3 we formally define the problem and how it relates to the heavy-traffic limits.

### 2.1 Heavy-traffic and State-Space Collapse

As we consider systems under heavy-traffic, we need to define several terms relevant in this setting.

**Definition 1** A random variable \( Y \) has an \((\alpha, \beta)\) impulse-exponential distribution if \( P(Y > 0) = \alpha \) and \( P(Y > x | Y > 0) = e^{-\beta x} \) for \( x > 0 \), for some \( 0 \leq \alpha \leq 1 \) and \( \beta > 0 \). A stochastic process \( Z \) is said to have an \((\alpha, \beta)\) impulse-exponential steady-state if it has a well defined stationary distribution \( Z(\infty) \) that is \((\alpha, \beta)\) impulse-exponential.

Consider a parametric family of queueing systems, indexed by \( n \), with each system fitting the description above where primitives are now superscripted by \( n \). Thus \( \rho^n \) is the traffic intensity of the \( n \)th system. We assume that the base system described in the previous section is associated with some index \( n^* \) so that \( \rho^{n^*} = \rho < 1 \). If \( \lim_{n \to \infty} \rho^n = 1 \) then the sequence of systems is said to be going into heavy-traffic. Heavy-traffic limit theorems consider the limiting processes of such a sequence of systems. We denote by “\(^-\)” the limit of any system primitives; therefore \( \bar{\rho}_i = \lim_{n \to \infty} \rho_i^n \), \( i = 1, 2 \) and \( \bar{\rho} = 1 \), by definition.
If \( \lim_{n \to \infty} \rho^n = 1 \) then queue lengths will be exploding and must be scaled for the limit to provide any sort of meaningful results. In particular, if \( X^n = \Psi(X, n) \) is some scaling of the number of customers in the system, then it is the limit of these scaled processes that are studied. This scaled limit process may then be unscaled to provide an approximation for system behavior in moderate traffic (i.e., \( X^n \approx \Psi^{-1}(\hat{X}, n) \), where \( \hat{X} \) is the limit of the processes \( X^n \)).

The convergence described in the previous paragraphs is weak convergence of stochastic processes and is usually defined either in terms of the Skorohod metric or in terms of uniform convergence on compact sets (see, e.g., Glynn 1990). Here, we merely assume that weak convergence, denoted by \( \Rightarrow \), is well defined given the sequence of systems under consideration. We prove our results under any heavy-traffic limit that results in an appropriate limit for scaled queue length. To wit, consider the following definition.

**Definition 2** Consider a sequence of systems fitting the assumptions above, operated under some policy \( \pi \), and indexed by \( n \) such that \( \rho^n \to 1 \). The number of customers in the system is scaled according to some deterministic function \( \Psi(\cdot, n) \) and let the scaled total queue length be represented by \( \hat{Q}^n \). We say that policy \( \pi \), together with this scaling and primitives, has an \((\alpha, \beta)\) impulse-exponential heavy-traffic limit if \( \hat{Q}^n \) converges weakly to a stochastic process \( \hat{Q} \) with \((\alpha, \beta)\) impulse-exponential steady-state.

We will prove our results under any general \((\alpha, \beta)\) impulse-exponential heavy-traffic limit.

The analysis of heavy-traffic systems is frequently aided by state-space collapse (see Reiman 1984 and Whitt 1971). The definition we give below is the one we believe to be most intuitive for the two-class system considered here and is in line with that in Mandelbaum and Stolyar (2004).

**Definition 3** We say that the heavy-traffic limit of a policy \( \pi \) exhibits state-space collapse if there exists a deterministic mapping \( u^\pi(\cdot) \) with \( 0 \leq u^\pi(\cdot) \leq 1 \) such that if \( \hat{Q}^n \Rightarrow \hat{Q} \) then \( \hat{Q}^n_i \Rightarrow \hat{Q}_i \) where for each \( t \geq 0 \), \( \hat{Q}_1(t) \overset{D}{=} u^\pi(\hat{Q}(t))\hat{Q}(t) \) and \( \hat{Q}_2(t) \overset{D}{=} (1 - u^\pi(\hat{Q}(t)))\hat{Q}(t) \). Further, there exist \( \lambda_i^\ast \) such that \( \hat{W}_i(t) \overset{D}{=} \hat{Q}_i(t)/\lambda_i^\ast \) for \( i = 1, 2 \).

State-space collapse in the context of this paper implies that, in the scaled heavy-traffic limit, one may probabilistically predict how many customers are in each queue, given the total number of customers in the system. For example, in priority queues all customers are in each queue, given the total number of customers in the system. Under FCFS service, if \( x \) is the total number of customers in the queue then each queue will have \( \hat{p}_i x \) customers. Note that we do not attempt to rigorously prove state-space collapse for our policies. Our optimal policy will be shown to be optimal among policies that exhibit state-space collapse. How broad a class this is, is left as an open research question.

The basic intuition behind state-space collapse is that queue lengths are moving on a fluid scaling, which is infinitely faster in the limit than the diffusion time scale. Thus, if on a fluid scale the system converges
to a fixed point within finite time, this time will be negligible on the diffusion scale. This fluid time scale is the basic idea used by Bramson (1998) and Williams (1998) to rigorously prove state-space collapse and the corresponding diffusion limits under conventional heavy-traffic scalings (described below). Also Mandelbaum, Massey and Reiman (1998) discuss the fluid approximation for a priority queue and Ridley, Massey and Fu (2004) provide numerical examples of such. The second part of the definition states that Little’s law will hold probabilistically for the actual queue lengths, instead of just on average. This again is a consequence of the deterministic relationship described above.

The service level agreements discussed in this paper consider averages for the system. As discussed in the introduction, this is a fundamentally different sort of constraint than that in most previous heavy-traffic work which requires the constraint be met exactly in heavy-traffic. Note that such an absolute requirement is only possible because of state-space collapse. Because we do not consider such absolute requirements, only averages, we must therefore consider a steady-state version of the problem.

**Definition 4** We say that the \((\alpha, \beta)\) impulse-exponential heavy-traffic limit from Definition 2, under state-space collapse from Definition 3, exhibits steady-state convergence if \(\tilde{Q}_n(\infty) \Rightarrow \hat{Q}(\infty)\), with \(\tilde{Q}^1_n(\infty) \Rightarrow u(\hat{Q}(\infty))\hat{Q}(\infty)\) and \(\tilde{Q}^2_n(\infty) \Rightarrow (1 - u(\hat{Q}(\infty)))\hat{Q}(\infty)\).

Steady-state convergence requires an interchange of limits that is rarely rigorously justified in the heavy-traffic limit. For example, Mandelbaum and Stolyar (2004) present Conjecture 1 which in effect hypothesizes steady-state convergence for their system. One exception to this is Garnett et al. (2002) who do prove the interchange (see their Theorem 2*). They are able to do this because they can explicitly calculate the stationary distributions of the sequence of scaled systems. It is not within the scope of this paper to explore when steady-state convergence occurs. Instead, we only consider systems where it does occur.

### 2.2 Examples

We consider how the definitions of the previous subsection apply under alternate heavy-traffic scalings. The “conventional” heavy-traffic sequence of systems has \(\sqrt{n}(\lambda^n - N\mu^n) \to a\), for some fixed constant \(a\) and number of servers \(N\), so that \(1 - \rho^n\) is approximately proportional to \(1/\sqrt{n}\). Here the heavy-traffic scaling \(\Psi(X^n, n)(t) = X^n(nt)/\sqrt{n}\). The limit for the scaled queue length process converges to a reflected (or regulated) Brownian motion with instantaneous drift \(a\), where \(a\) is the limit described above, and instantaneous variance:

\[ \dot{\sigma}^2 := \hat{\lambda}(\dot{c}_a^2 + \dot{c}_s^2), \]

(see Iglehart and Whitt 1970). This limit holds regardless of policy considered so long as the system is non-anticipatory and work conserving. If \(a < 1\) then this diffusion has an \((\alpha, \beta)\) impulse-exponential steady-state
with $\alpha = 1$ and $\beta = -2a/\hat{\sigma}^2$. As stated earlier, Bramson (1998) and Williams (1998) prove state-space collapse in this setting. Here $\tilde{W}_i(t) \overset{D}{=} \tilde{Q}_i(t)/(\hat{p}_i \hat{\mu} N^n)$ for $i = 1, 2$.

Halfin and Whitt (1981) introduced a many server heavy-traffic regime, where $n$ represents the number of servers (i.e., $N^n = n$ and $N^n^* = N$) and $\lim_{n \to \infty} \sqrt{n}(1 - \rho^n) = \beta$, for some constant $\beta$; thus $1 - \rho^n$ is again approximately proportional to $1/\sqrt{n}$. Under this limit, the probability of delay lies between zero and one. Under Halfin-Whitt scalings $\Psi(X^n, n)(t) = (X^n(t) - n)/\sqrt{n}$, $t \geq 0$. For an $M/M/N$ system, letting $\rho^n := \lambda^n/n\mu$, as $n \to \infty$, $\rho^n = 1 - \beta/\sqrt{n} + o(1/\sqrt{n})$. If $X^n(0) \Rightarrow \hat{X}(0)$, $X^n \Rightarrow \hat{X}$ where $\hat{X}$ is a one dimensional diffusion process with infinitesimal drift:

$$a(x) = \begin{cases} -\mu\beta & x \geq 0 \\ -\mu(x + \beta) & x < 0 \end{cases}$$

and infinitesimal variance $2\mu$. That is, the limiting distribution is two Ornstein-Uhlenbeck processes with different drifts centered at 0. Letting $\hat{Q}^n = (\hat{X}^n)^+$ and $\hat{Q} = \lim_{n \to \infty} \hat{Q}^n$, the steady-state distribution of the limiting queue length process has

$$P(\hat{Q}(\infty) > x|\hat{Q}(\infty) > 0) = e^{-\beta x}$$

and the limiting probability that a customer waits is given by

$$P(\hat{Q}(\infty) > 0) = \alpha \in (0, 1).$$

where

$$\alpha \equiv \alpha(\beta) = (1 + \sqrt{2\pi}/\beta \Phi(\beta)e^{\beta^2/2})^{-1},$$

$\beta$ is the limit given above, and $\Phi(x)$ is the standard Normal cumulative distribution function (CDF). Thus, the scaled queue length process has an $(\alpha, \beta)$ impulse-exponential steady-state. For the case of a $G/GI/N$ queue, Whitt (2004) approximates the queue length process as an impulse-exponential process with

$$P(\hat{Q}(\infty) > 0) = \alpha(\beta/\sqrt{z})$$

where $z = 1 + (c_a^2 - 1)\eta(G)$, $G$ is the service-time CDF, assumed to have finite mean, $G^c \equiv 1 - G$ is the associated complementary CDF, and

$$\eta(G) \equiv \mu \int_0^\infty G^c(x)^2 dx.$$

Also,

$$P(\hat{Q}(\infty) > x|\hat{Q}(\infty) > 0) = e^{-(\beta/v)x}$$

where $v = (c_a^2 + w^2 c_a^2 + 1 - w)/2$ for some value of $w$ decreasing in $\beta$ with $w(\beta = 0) = 1$ and $w(\beta = \infty) = 0$. Whitt suggests letting $w = 1$ though the approximation may improve for other values.
With respect to Definition 3, if state-space collapse exists for the Halfin-Whitt regime, $\lambda^*_i = \hat{p}_i \hat{\mu}_i$, i.e., $\hat{W}_i(t) \overset{D}{=} Q_i(t)/(\hat{p}_i \hat{\mu}_i)$ for $i = 1, 2$. We note though that state-space collapse for systems under Halfin-Whitt scalings is an active and interesting open research area. Atar et al. (2004b) state that: “While such [state-space] collapse prevails in the special case studied in [Atar et al. (2004a)], simulations and intuition indicate that, in general for the [Halfin-Whitt] regime, an analogous phenomenon is unlikely to occur.” State-space collapse under the Halfin-Whitt regime is also proven for specific scheduling policies in Armony (2003), Armony and Maglaras (2004a, 2004b), and Gurvich (2004). However, Harrison and Zeevi (2004) present a situation where it is absent. As some of our examples below utilize the Halfin-Whitt regime, we should caution that state-space collapse may not, in fact, occur in reality for these examples.

2.3 System Objectives

In this paper we consider the response of a firm to alternate contracts, and so the system objectives of the firm depend on the form of these contracts and the firm’s objective with respect to non-contract customers. Suppose that contract customers are guaranteed a minimum average performance on some measure of delay. Let $\Pi$ be the set of all Head-of-Line policies. Let $g^\pi(W_1)$ be the random variable of interest for a given policy $\pi \in \Pi$. For example, under a delay percentile contract, for a given policy $\pi$, $g^\pi(W_1) = I_{W_1 \leq D}$ and the contract requires $E[g^\pi(W_1)] \geq \gamma$. Non-contract customers have no guarantees on delay. As discussed in the introduction, the firm wishes to give them good delay performance in order to increase the chance that they repeat their business or even convert to contract customers. We do not explicitly model either of these motivations, and instead assume we seek to minimize the average of some measure of the non-contract customers’ delay, $h^\pi(W_2)$. Such was the approach of the call center that served as a motivation for this research. For simplicity, throughout the paper we assume that $h^\pi(W_2) = W_2$. The problem we consider is:

$$\min_{\pi \in \Pi} E[h^\pi(W_2)] \text{ s.t. } E[g^\pi(W_1)] \geq K,$$

for some constant $K$.

As discussed earlier, in order to approach this problem, we consider an appropriately scaled liming version of the problem in heavy traffic. Also as noted earlier, the form of both the objective and constraint differs from those traditionally used in the heavy-traffic literature because they consider averages. For these reasons we consider asymptotic optimality among policies and sequences of scaled systems that:

1. Have an $(\alpha, \beta)$ impulse-exponential heavy-traffic limit;
2. Exhibit state-space collapse; and
3. Exhibit steady-state convergence.
While such a definition of asymptotic optimality is in some sense more restrictive than those in earlier heavy-traffic work because it carries with it a number of assumptions it is also more general because it provides results for system averages.

In order to define the limiting problem, let \( \hat{g}^\pi(\cdot) \) and \( \hat{h}^\pi(\cdot) \) be appropriately scaled functions of interest such that:

\[
\hat{g}^\pi(\tilde{W}_n^1) = g^\pi(W_n^1), \quad \hat{h}^\pi(\tilde{W}_n^2) = h^\pi(W_n^2),
\]

\[
\hat{g}^\pi(\hat{W}_1) = \lim_{n \to \infty} \hat{g}^\pi(\tilde{W}_n^1)
\]

and

\[
\hat{h}^\pi(\hat{W}_2) = \lim_{n \to \infty} \hat{h}^\pi(\tilde{W}_n^2).
\]

Note that sufficient conditions for the above convergence are that \( \hat{g}^\pi \) and \( \hat{h}^\pi \) satisfy the conditions of the continuous mapping theorem (see, e.g., Glynn 1990). Then, by our assumption of state-space collapse, we can express \( \hat{W}_i \overset{D}{=} \hat{Q}_i/\lambda_i^* \) and the policy \( \pi \) through \( u(\hat{Q}) \) to arrive at the limiting problem:

\[
\min_{0 \leq u(x) \leq 1} \quad E_{\hat{Q}} \left[ \hat{h} \left( \frac{1}{\lambda_2^*} \hat{Q}(1 - u(\hat{Q})) \right) \right] \\
\text{s.t.} \quad E_{\hat{Q}} \left[ \hat{g} \left( \frac{1}{\lambda_1^*} \hat{Q} u(\hat{Q}) \right) \right] \geq K. \tag{1}
\]

If we find a \( u(\cdot) \) that solves (1) then the task becomes one of finding a real-time policy that converges to this limit. Having done so, such a policy is then asymptotically optimal among all systems with an impulse-exponential heavy-traffic limit which exhibit state-space collapse and steady-state convergence. As mentioned earlier, we leave investigation of the size of this class as the subject of future research.

3 Main Results

This section gives the main results of the paper. Section 3.1 derives the asymptotically optimal policy for the firm operating under a delay percentile contract. It also presents a regular-traffic policy designed to result in this asymptotic policy. Section 3.2 analyzes the performance of the policy found in Section 3.1, both through asymptotic analysis and through simulations, and discusses the practical implications of this performance. Finally, Section 3.3 presents some new contract forms that may be more appealing from a contract customer’s perspective, analyzes their performance, and discusses practical implications of the work.

3.1 Delay Percentile Contracts

In this subsection we assume that contract customers are operating under a delay percentile contract. We derive the optimal policy for the firm, under the assumptions of heavy-traffic steady-state, as described in
the previous section. We then present a regular-traffic policy that we argue should asymptotically converge to the optimal heavy-traffic policy. The implications of these results are discussed.

Under the previously described assumptions, in the case of a delay percentile contract, 

\[ \hat{h}(w) = w \]

and 

\[ \hat{g}(w) = I_{w \leq \hat{D}}, \]

where \( \hat{D} \) is the acceptable delay in the contract scaled appropriately so that \( \hat{g}^\alpha(W_1^\alpha) = g^\alpha(W_1^\alpha) \). Thus, under conventional heavy-traffic scalings \( \hat{D} = D/\sqrt{n} \). Under Halfin-Whitt scalings \( \hat{D} = \sqrt{N \hat{D}} \), where \( N = N_n^* \) is the number of servers in the non-scaled system when the delay \( D \) is contracted. To see this, note that under Halfin-Whitt, while the queue length grows as \( \sqrt{n} \), the time for \( n \) servers to serve these customers decreases as \( 1/\sqrt{n} \). Under conventional scalings both the queue length and delay grow in proportion to \( \sqrt{n} \). Therefore, the given scalings for contracted delay provide the appropriate scaled delays for the corresponding systems.

By definition, \( \hat{h}(\tilde{W}_2^\alpha) = h^\alpha(W_2^\alpha) \). Further, 

\[ \hat{h}(\tilde{W}_2) \overset{D}{=} \hat{h}(\hat{Q}_2/\lambda_2) = \hat{Q}_2/\lambda_2 \]

and 

\[ \hat{g}(\tilde{W}_1) \overset{D}{=} \hat{g}(\hat{Q}_1/\lambda_1^*) = I_{\hat{Q}_1/\lambda_1^* \leq \hat{D}}. \]

Observe that both \( \hat{g} \) and \( \hat{h} \) are continuous except for \( \hat{g} \) at \( \hat{D} \). So long as \( \hat{D} > 0 \) and \( \hat{Q}_1 \) is impulse-exponential (and therefore has zero mass at \( \hat{D} \)), both \( \hat{g} \) and \( \hat{h} \) satisfy the conditions of the continuous mapping theorem. The optimization is therefore:

\[
\min_{0 \leq u(x) \leq 1} \mathbb{E}_Q \left[ \frac{1}{\lambda_2} x (1 - u(x)) \right] \\
\text{s.t.} \\
\mathbb{E}_Q \left[ I_{x u(x) \leq \lambda_1^* \hat{D}} \right] \geq \gamma.
\]

Let \( F_{\hat{Q}}(\cdot) \), be the CDF for \( \hat{Q} \), which is \( (\alpha, \beta) \) impulse-exponential by assumption. Further, define \( \hat{Q}_L = \lambda_1^* \hat{D} \). The intuition behind \( \hat{Q}_L \), which will be formally shown in the below proposition, is that for total queue lengths below \( \hat{Q}_L \), even if all customers are found in queue 1, the contract can still be met asymptotically. A similar intuition is used in Armony and Maglaras (2004b). Note that all of the proofs of propositions and observations (smaller results) may be found in the appendix.

**Proposition 1** The optimal control for the delay percentile contract problem given in (2), is \( u(x) = 1 \) for \( x \leq \hat{Q}_L \) and \( x \geq \hat{Q}_H \) and \( u(x) = \hat{Q}_L/x \) for \( \hat{Q}_L < x < \hat{Q}_H \), where \( \hat{Q}_H \) satisfies \( F_{\hat{Q}}(\hat{Q}_H) = \gamma \).
We refer to the policy in Proposition 1 as a Two-Point Threshold Policy. We observe that in the heavy traffic limit, the non-contract customers are given priority either when there is relatively little demand, so that even by delaying the contract customers their contract terms are met, or when there is a high level of demand, so that the cost of delaying the non-contract customers is reduced. Observe that, if one considers the queue length of the contract customers in scaled steady-state heavy-traffic,

\[
\hat{Q}_1(t) = \begin{cases} 
\hat{Q}(t) & \text{if } \hat{Q}(t) \leq \hat{Q}_L \\
\hat{Q}_L & \text{if } \hat{Q}_L < \hat{Q}(t) < \hat{Q}_H \\
\hat{Q}(t) & \text{if } \hat{Q}(t) \geq \hat{Q}_H 
\end{cases}
\] (3)

That is, in scaled heavy traffic, the queue length for class 1 customers is allowed to build to the length of \(\hat{Q}_L\). As the queue continues to build, the additional queueing customers are from class 2. When the total queue length exceeds \(\hat{Q}_H\), which occurs \((1 - \gamma) \times 100\%\) of the time, class 2 customers are given priority to reduce the cost.

The optimal policy differs from the threshold rule in Armony and Maglaras (2004a) in that they require only a single threshold value, namely the lower one \(Q_L\). In their problem, the contract requires in the limit as \(n \to \infty\), \(\hat{W}_1(t) \leq \hat{D}(t)\) for all \(t > 0\) (after some notational changes). As the delay percentile contract has two parameters, we find two thresholds, one corresponding to the delay required and one corresponding to the percentile of interest.

**Regular Traffic Policies:** The policy above is well defined under heavy-traffic systems with state-space collapse; however, we must consider what a regular (non-heavy) traffic policy that yields this optimal policy asymptotically would look like.

To this end, let \(Q_L = \lambda_1 D\) and let \(Q_H = \sqrt{n} \cdot \hat{Q}_H\) be \(\hat{Q}_H\) unscaled so that \(\text{Prob}(Q < Q_H) \approx \text{Prob}(\hat{Q} < \hat{Q}_H) = \gamma\). Let \(\pi^A\) be the following non-preemptive, work-conserving policy. Each class is served in FCFS order. When a server becomes available if \(Q_L < Q_1(t)\) and \(Q(t) < Q_H\) then class 1 customers are served if available, otherwise class 2 customers (if any) are served. However, if when a server becomes available either \(Q_1(t) \leq Q_L\), \(Q(t) \leq Q_L\), or \(Q(t) \geq Q_H\) then class 2 customers are served, if possible, otherwise class 1 customers (if any) are routed to the free server.

We conjecture that a sequence of systems operated under \(\pi^A\) and scaled appropriately will converge to the heavy-traffic system and policy given above. In the on-line appendix we prove that in a fluid system operated under our policy the queue lengths indeed converge to a fixed point in finite time; such a proof forms the core of most of the state-space collapse results previously given and thus provides strong evidence for the hypothesized convergence.
3.2 Performance Analysis

In this section we observe that the asymptotic delay of the class 1 customers under the Two-Point Threshold policy may be (1) highly variable and (2) significantly greater than the non-contract customers.

To this end, we observe, if $\hat{Q}_L \leq \hat{Q}_H$,

$$E[\hat{Q}_1] = \int_0^{\hat{Q}_L} xdF_{\hat{Q}}(x) + \int_{\hat{Q}_L}^{\hat{Q}_H} \hat{Q}_L dF_{\hat{Q}}(x) + \int_{\hat{Q}_H}^{\infty} xdF_{\hat{Q}}(x)$$

$$= \frac{\alpha}{\beta} e^{-\hat{Q}_H} \left( e^{\hat{Q}_L} - e^{\hat{Q}_L + \hat{Q}_L} + e^{\hat{Q}_H + \hat{Q}_L} + e^{\hat{Q}_L} \left( \hat{Q}_H \beta - \hat{Q}_L \beta \right) \right)$$

and

$$Var[\hat{Q}_1] = \frac{\alpha}{\beta^2} e^{-\hat{Q}_L} \left( e^{\hat{Q}_H + \hat{Q}_L} - 2 e^{\hat{Q}_H} + \left( 1 + \hat{Q}_L \beta \right) + e^{\hat{Q}_L} \left( 2 + 2 \hat{Q}_H \beta + \hat{Q}_H \beta^2 - \hat{Q}_L \beta^2 \right) \right);$$

and if $\hat{Q}_H < \hat{Q}_L$,

$$E[\hat{Q}_1] = \alpha/\beta \text{ and } Var[\hat{Q}_1] = (2 - \alpha) \alpha/\beta^2.$$  

Also we find for $\hat{Q}_H \geq \hat{Q}_L$

$$E[\hat{Q}_2] = \alpha/\beta e^{-\beta(\hat{Q}_L + \hat{Q}_H)} \left( e^{\beta \hat{Q}_H} - e^{\beta \hat{Q}_L} - \beta e^{\beta \hat{Q}_L}(\hat{Q}_H - \hat{Q}_L) \right)$$

and $E[\hat{Q}_2] = 0$ otherwise.

In Table 1 we present numerical results from the heavy-traffic approximation and from simulation for several contracts terms ($\gamma, D$) and demand scenarios (changing $p_i$) while holding $N = 50, \mu = 1$ and $\Lambda = 49$. In all examples the arrivals are Poisson and service times are exponentially distributed. We actually simulated a minor variant of the regular-time policy proposed in Section 3.1. This was because, while asymptotically one can guarantee that delay constraints can be met even if queue 1 is given lower priority for $Q_1(t) < Q_L$, in practice we found that such a lower priority does not work well for regular traffic. The simulated policy therefore only gives priority to class 2 if $Q(t) \leq Q_L$ or if $Q(t) \geq Q_H$, where $Q_L$ and $Q_H$ were determined under a Halfin-Whitt heavy-traffic scaling. The system was simulated for 2,000,000 time units, where the first 10% of data are deleted in order to avoid problems with initial transience. Common random numbers are used across the runs to make the comparisons more accurate.

In the table $\gamma^{SIM}$ is the observed value of $\gamma$ and $Q_H^{SIM}$ is the value of $Q_H$ required for the contract $\gamma$ to be met exactly (holding $Q_L$ constant). We observe that in most cases the simulated system almost met the contract and by increasing $Q_H$, the contract could be met. We observe that case 4 and case 8 had the largest deviations from the heavy-traffic approximation. Case 4 had the most restrictive contract ($\gamma = 0.95$)
Table 1: Results for heavy-traffic (Halfin-Whitt regime) and simulation study. For all cases $N = 50$, $\mu = 1$, $\Lambda = 49$ resulting in $\alpha = 0.8339$ and $\beta = \sqrt{2}/10 \approx 0.1414$.

while in Case 8, the contract customers dominate ($p_1 = 0.8$). One would expect the approximation to be challenged in these circumstances.

We also present the expected delay for class 1 and class 2 under the heavy-traffic approximation (H-T) and for the simulation (SIM). The observed delays are reasonably close to those of the H-T approximation. In the table we observe that depending on the case, the expected delay for class 1 may be greater than the expected delay for class 2 customers (both in the approximation and in the simulation). In fact, we can show the following:

**Observation 1** If

$$\lambda_2 \text{Exp}[\beta(Q_H + Q_L)] - (\lambda_1 + \lambda_2)(\text{Exp}[\beta Q_H] - \text{Exp}[\beta Q_L]) + \beta(\lambda_1 + \lambda_2)\text{Exp}[\beta Q_L](Q_H - Q_L) > 0,$$

the approximation implies $E[\text{Delay for Class 1}] > E[\text{Delay for Class 2}]$.

The observation and the numerical examples indicate several problems with the delay percentile contract. Following intuition, the observation implies (and the numerical results confirm) that as the contract becomes less restrictive, either by allowing longer delays or lower percentiles, class 1 is disadvantaged relative to class 2. Similarly, class 1 has worse performance when they are in the minority. Significantly, we observe that for the so called ‘80/20’ contracts (80% of customers served within 20 seconds or 0.333 minutes) the contract
customer’s average delay is greater than that of the non-contract customers (except when contract customers were in the vast majority).

In Table 1, we also present the coefficient of variation (CV) for the scaled class 1 queue length, the CV for the actual wait times for class 1 customers, and the ratio of the CVs for the scaled queue length to that if the call center used a FCFS policy for all customers. Under FCFS service the CV for the scaled queue length is $2/\alpha - 1$. We observe that the coefficient of variation for the limiting class 1 queue length is comparable to the simulated CV of the waiting time as would be expected under our system. Further, both are large compared with the coefficient of variation for a FCFS system. The high values for the contract customers indicates that they observe heavy tails in their service distribution. This is clearly the effect of the firm’s policy not to give them priority during periods of excessive delays.

Thus, the delay percentile contract has the unappealing properties that delays may be long (relative to both other customers and what might be achieved without a contract) and CVs can be high compared to a nominal value given by a FCFS policy. High variability or heavy tails is almost certainly not what the contract customers had in mind when signing a delay percentile contract. Of course, if the contract customers can persuade the firm to offer 100% guarantees on the delay then the undesirable behavior exhibited above is mitigated. We turn next to investigating alternate contracts in order to construct some recommendations to address the above problems for the contract customers.

3.3 Alternate Contracts

In this subsection we examine several alternate contracts and analyze the optimal heavy-traffic policy for the firm operating the contract. Here, rather than solely contracting on a percentile of delay, we consider contracts where penalties are incurred based on some measure of delay. We show that several simple penalty contracts do not lead to good performance, either for the contract customers or for reasonable expectations for the system as a whole. However, we show that by combining a percentile delay contract with a convex penalty function, the mean delay for contract customers and its CV decrease substantially. However, as this diminishes the service vis-a-vis non-contract customers, we also investigate contract structures that may reduce the inequality between them.

First, consider a contract for which the firm incurs a fixed cost for each call whose queue time exceeds a given delay $D$. If the firm minimizes this cost plus a linear delay cost for non-contract customers, the problem is simply:

$$\min_{\pi \in \Pi} c_1 E[1_{W_1 \geq D}] + c_2 E[W_2]$$
and the limiting heavy-traffic problem is

\[ \min_{0 \leq u(x) \leq 1} c_1 E_Q[1_{xu(x) > Q_L}] + c_2 E_Q[x(1 - u(x))] \]  \hspace{1cm} (4) \]

If one considers the value of \( c_1 \) as a dual variable on the constraint of the Delay Percentile Contract (2), by choosing \( c_1 \) appropriately, (4) can be made equivalent to (2). Similarly, by choosing \( c_2 \) appropriately, the contract can be made equivalent to one where the firm chooses to minimize the cost for excessive delay of contract customers subject to a constraint on the mean delay incurred by the non-contract customers. That is, if we consider the problem:

\[ \min_{\pi \in \Pi} c_1 E[1_{W_1 > D}] \]
\[ \text{s.t.} \]
\[ E[W_2] \leq K \]

for some \( c_1 \) and \( K \), the limiting heavy-traffic problem is

\[ \min_{0 \leq u(x) \leq 1} c_1 E_Q[1_{xu(x) > Q_L}] \]
\[ \text{s.t.} \]
\[ E_Q[x(1 - u(x))] \leq K' \] \hspace{1cm} (5) \]

for an appropriately scaled \( K' \).

**Observation 2** The fixed cost problems (4) and (5) are equivalent to the delay percentile contract.

Second, consider a contract for which the cost the firm incurs is linear in the delay of the contract customers. If the firm minimizes this cost subject to a constraint on the mean delay of non-contract customers, then in the limiting problem in heavy traffic, we observe that all solutions that delay the non-contract customers as much as possible are optimal. That is, consider the implied limiting problem

\[ \min_{\pi \in \Pi} cE[W_1] \text{ s.t. } E[W_2] \leq K, \]

the limiting heavy-traffic problem (stated in terms of the scaled queue length) is:

\[ \min_{0 \leq u(x) \leq 1} c_1 E_Q[xu(x)] \]
\[ \text{s.t.} \]
\[ E_Q[x(1 - u(x))] \leq K' \] \hspace{1cm} (6) \]

for an appropriate value of \( K' \).

**Observation 3** All policies for which the constraint is tight are optimal for the linear cost problem (6).

Thus dismissing with the fixed cost and linear cost contracts, we consider convex cost contracts.
Convex Cost Contract  As discussed in the introduction, call center practitioners advise that call center contracts impose increasing penalties for delaying customers. As an initial consideration, suppose that a “convex cost contract” imposes a cost to the firm for delaying contract customers that is a strictly convex increasing measure of the delay. Such a contract would be a relatively simple means of enforcing an increasing penalty for delay. Let $\phi(\cdot)$ be the appropriately scaled cost function. If the firm minimizes its cost subject to a constraint on the mean delay of non-contract customer, the limiting problem is

$$\min_{0 \leq u(x) \leq 1} E_Q[\phi(xu(x))]$$

s.t.

$$E_Q[x(1-u(x))] = K'$$

(7)

**Proposition 2** The solution to the limiting convex cost problem (7) gives priority to class 2 customers if $\hat{Q}(t) < \hat{Q}^*$ and priority to class 1 customers otherwise, where $\hat{Q}^*$ solves $\int_0^{\hat{Q}^*} xd\hat{F}_Q(x) = K'$. As in the case of the fixed cost problem, we can show that rather than requiring $E[W_2] = K$, a formulation minimizing $E[\phi(W_1)] + c_2E[W_2]$ would give the same solution (for an appropriately chosen value of $c_2$).

Letting $\hat{Q}_L = \lambda_1 \hat{D}$ as in Section 3.1, under the Convex Cost contract, if $\hat{Q}^* > \hat{Q}_L$, $P(\hat{W}_1 > \hat{D}) = \int_{\hat{Q}_L}^{\hat{Q}^*} d\hat{F}_Q(x) = \hat{F}_Q(\hat{Q}^*) - \hat{F}_Q(\hat{Q}_L)$. Thus by appropriately choosing $\hat{Q}^*$ or equivalently, $K$, the firm can achieve a given percentile delay $\gamma$ in such a contract as well. Recall that the delay percentile contract considered in Section 3.1 was determined by minimizing $E[\hat{W}_2]$. Let $K^* = E[\hat{W}_2]$ for the optimal two-threshold policy. Then $K \geq K^*$ if the optimal policy for the Convex Cost contract is to achieve the same percentile $\gamma$. Observe that for the Convex Cost contract, rather than the queue length increasing according to (3), the queue would have the form

$$\hat{Q}_1(t) = \begin{cases} 
\hat{Q}(t) & \text{if } \hat{Q}(t) \leq \hat{Q}^* \\
0 & \text{otherwise.} 
\end{cases}$$

The Convex Cost contract has the appealing property (from the contract customers’ viewpoint) that (at least in the scaled limit) contract customers never see the tail of the distribution. When queues are large, it is the non-contract customers that see the delays. This contrasts with the percentile-delay contract where it is the contract customers who observe the large delays. Let $\hat{Q}_1^{PD}$ be the random variable for class 1 queue length under the optimal policy for the Delay Percentile contract and let $\hat{Q}_1^{CC}$ be that for the Convex Cost contract. We have the following:

**Observation 4** If $P(\hat{Q}_1^{PD} \leq \hat{Q}) = P(\hat{Q}_1^{CC} \leq \hat{Q}) = \gamma$, $E[Q_1^{CC}] < E[Q_1^{PD}]$.

As in the case for the Delay Percentile contract, we can formulate a regular traffic policy that should yield the optimal policy asymptotically. Let $Q_L = \lambda_1 D$ as before and let $Q^* = \sqrt{n^*} \hat{Q}$ so that $\text{Pr}(Q_L < Q <


Table 2: Mean Delay for Class 1 and Class 2, observed $\gamma_{SIM}$ and $CV(W_1)$ from simulation for the Delay Percentile, Convex Cost, Combined and Delay Difference-FCFS contracts for the 12 cases presented in Table 1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E[W_1]$</td>
<td>$E[W_2]$</td>
<td>$\gamma_{SIM}$</td>
<td>$CV(W_1)$</td>
</tr>
<tr>
<td>1</td>
<td>1.027</td>
<td>0.688</td>
<td>0.773</td>
<td>2.087</td>
</tr>
<tr>
<td>2</td>
<td>0.876</td>
<td>0.838</td>
<td>0.810</td>
<td>2.389</td>
</tr>
<tr>
<td>3</td>
<td>0.684</td>
<td>1.031</td>
<td>0.853</td>
<td>2.893</td>
</tr>
<tr>
<td>4</td>
<td>0.612</td>
<td>1.102</td>
<td>0.898</td>
<td>3.134</td>
</tr>
<tr>
<td>5</td>
<td>2.186</td>
<td>0.525</td>
<td>0.760</td>
<td>2.360</td>
</tr>
<tr>
<td>6</td>
<td>1.243</td>
<td>0.600</td>
<td>0.773</td>
<td>2.149</td>
</tr>
<tr>
<td>7</td>
<td>0.893</td>
<td>0.804</td>
<td>0.766</td>
<td>2.004</td>
</tr>
<tr>
<td>8</td>
<td>0.759</td>
<td>1.251</td>
<td>0.720</td>
<td>1.751</td>
</tr>
<tr>
<td>9</td>
<td>2.206</td>
<td>0.520</td>
<td>0.693</td>
<td>2.335</td>
</tr>
<tr>
<td>10</td>
<td>1.314</td>
<td>0.553</td>
<td>0.748</td>
<td>2.006</td>
</tr>
<tr>
<td>11</td>
<td>1.035</td>
<td>0.591</td>
<td>0.774</td>
<td>1.681</td>
</tr>
<tr>
<td>12</td>
<td>0.935</td>
<td>0.548</td>
<td>0.765</td>
<td>1.364</td>
</tr>
</tbody>
</table>

$Q^* \approx 1 - \gamma$. Let $\pi^{CC}$ be the following non-preemptive, work-conserving policy. Each class is served in FCFS order. When a server becomes available and $Q(t) < Q^*$ then class 2 customers are served if available, otherwise class 1 customers (if any) are served. However, if when a server becomes available $Q(t) \geq Q^*$ then class 1 customers are served if possible, otherwise class 2 customers (if any) are routed to the server. We next compare the performance of the Convex Cost policy to that of the Percentile Delay contract.

In Table 2, we present the mean delay for both customer classes as well as the fraction served within the contracted delay time and the CV of the observed delay for class 1 customers provided by simulating the system using the value of $Q^*$. We do so for the Delay Percentile contract and the Convex Cost contract (as well as the Combined and Delay Difference-FCFS contracts discussed below). In accordance with Proposition 4, we observe that the mean delay for class 1 customers under the Convex Cost contract is reduced vis-a-vis the Delay Percentile contract. We also observe that the CV of the Convex Cost contract is also reduced, though not by the same order of magnitude as the mean. We note that in the heavy-traffic the CV for the scaled queue length for the Delay Percentile contract is not necessarily smaller than that for the Convex Cost contract. (Though not shown, we note that the mean delay and CV values presented are close to their expected values given by the H-T approximation. Differences are on the same order as those shown in Table 1.)

However, the reduced delay may not be as appealing as one might initially think. In particular, if the cost of a contract is related to the expected delays imposed by contract customers on non-contract customers, one
might expect that having zero delay when the queue is long ($\hat{Q} > \hat{Q}^*$), would be an unnecessary cost for the contract customers. Rather, the delay $D$ from the Delay Percentile contract would seem to be acceptable in this case as one could argue that such would be the expected delay of contract customers when the system is busy.

**Combined Percentile-Delay/Convex Cost Contract** In order to address this problem, consider a Combined contract where a fraction $\gamma$ of all customers wait less than $D$ time units and the firm is charged a convex increasing cost, $\phi$, for all delays in excess of $D$. If the firm minimizes this penalty cost plus a linear cost for the expected delay of class 2 customers, the problem is then:

$$\min_{x \in \Pi} E[\phi((W_1 - D)^+) + c_2E[W_2]$$

$$\text{s.t. } E[1_{W_1 \leq D}] \geq \gamma$$

for which the limiting problem is:

$$\min_{0 \leq u(x) \leq 1} \quad E_\hat{Q}[\phi(xu(x))] + c'_2E_\hat{Q}[x(1 - u(x))]$$

$$\text{s.t. } E_\hat{Q}\left[\frac{1}{1 \cdot xu(x)}\right] \geq \gamma.$$ (8)

We have the following:

**Proposition 3** The optimal solution for the limiting problem for the Combined Contract (8) is $u(x) = 1$ for $x \leq \hat{Q}^*$ and $u(x) = \hat{Q}_L/x$ otherwise, where $F_{\hat{Q}}(\hat{Q}^*) - F_{\hat{Q}}(\hat{Q}_L) = \gamma$.

We observe that the optimal solution to Combined contract gives priority to non-contract customers until $(1 - \gamma) \times 100\%$ of the class 1 customers have delays in excess of $\hat{D}$. After that point, the contract customers are delayed until their delay equals $\hat{D}$. The regular traffic policy that we hypothesize should converge to the heavy-traffic optimal policy is as follows: Class 2 customers are given priority if $Q(t) \leq Q^*$ or if $Q(t) > Q^*$ and $Q_1(t) \leq Q_L$; otherwise class 1 customers are given priority.

Letting $\hat{Q}_1^{CB}$ be the random variable for class 1 queue length under the optimal policy for the Combined Contract, we have:

**Observation 5** Under the condition that the Percentile Delay, Convex Cost and Combined contracts use a common $\gamma$, then $E[\hat{Q}_1^{CC}] \leq E[\hat{Q}_1^{CB}] \leq E[\hat{Q}_1^{PD}]$

In Table 2, we observe that the mean delay for class 1 customers under the combined contract is on average 36% of the delay for the Delay Percentile contract. Further the CV is reduced on average by 55%. The implication is that by observing both the Delay Percentile contract as well as ensuring that the delays are
not too long, contract customers can reduce both the mean and variance of their delays. However, the mean class 1 delay is higher than that for the convex cost contract, as here the delay is $\hat{D}$ rather than 0 for $\hat{Q} > \hat{Q}^*$. The result is that the average over the test cases of the mean delay is 89% of the contracted delay $D$ rather than only 41%. This is because the non-contract customers receive comparatively better service under this contract when the system is busy.

**Delay Percentile Contract with a Delay Difference Penalty**  A more sophisticated buyer of a service level contract might see the problems inherent in the delay percentile contract, namely that for high demand scenarios the firm might give low priority to contract customers since the contract may be fulfilled without prioritizing these customers. Further, the buyer might also see that in cases of low demand, the contract customers would receive low priority. Recognizing these problems, as a potential solution to the problem, the buyer might want a contract that leads to FCFS service when contract customers do not receive priority. We show that one way to achieve this is to include a penalty for any negative treatment of contract customers as compared to non-contract customers.

We therefore consider a Delay Percentile Contract that imposes a penalty term when the non-contract customers receive better service than contract customers. Specifically, suppose the penalty reflects the squared difference between $W_2$ and $W_1$ when it is positive. (For simplicity we assume a square term as opposed to a general convex function; the results would be similar.) This contract is more complex than the previous ones considered as it requires reporting on the wait times for both contract and non-contract customers. We include it here for comparative purposes, noting actual implementation and monitoring would be difficult.

The problem for a firm that seeks to minimize the wait times of non-contract customers under a Delay Difference contract is:

$$\min_{\pi \in \Pi} \quad c_1 E[(W_1 - W_2)^+] + c_2 E[W_2]$$

s.t.

$$E[1_{W_1 \leq D}] \geq \gamma$$

The limiting problem for the firm under a Delay Difference contract is:

$$\min_{0 \leq u(x) \leq 1} \quad c_1 E_{\hat{Q}} \left[ \left( \frac{x u(x)}{\lambda_1} - \frac{x (1-u(x))}{\lambda_2} \right)^2 \right] + c_2 E_{\hat{Q}} \left[ \frac{x (1-u(x))}{\lambda_2} \right]$$

s.t.

$$E_{\hat{Q}} \left[ \frac{1}{\lambda_1} x u(x) \leq \hat{D} \right] \geq \gamma.$$  

(9)

We have the following

**Proposition 4** Let $a = \frac{c_2 (\lambda_1^* + \lambda_2^*)^2}{2 \sigma_1 (\lambda_1^* + \lambda_2^*)}$ and let $\hat{Q}_M = \frac{\hat{Q}_L (\lambda_1^* + \lambda_2^*)}{\lambda_1^*} - \frac{c_2 \lambda_1^* \lambda_2^*}{2 \sigma_1 (\lambda_1^* + \lambda_2^*)}$ where $\hat{Q}_L = \lambda_1^* \hat{D}$ and define
\[ p^*_1 = \frac{\lambda_1^*}{(\lambda_1^* + \lambda_2^*)}. \] The optimal policy for the limiting problem is

\[
u(x) = \begin{cases} 
\frac{a}{x} + p^*_1 & \text{if } x \leq \hat{Q}_M \\
\frac{\hat{Q}_L}{x} & \text{if } \hat{Q}_M < x \leq \hat{Q}_H \\
\frac{a}{x} + p^*_1 & \text{if } x > \hat{Q}_H
\end{cases}
\]

Observe that \( \lim_{c_1 \to -\infty} a = 0 \) and \( \lim_{c_1 \to -\infty} \hat{Q}_M = (\lambda_1^* + \lambda_2^*) \hat{D} \), so that the optimal policy for the Delay Difference contract in the limit has FCFS service if \( x \leq \hat{D} \) or \( x \geq \hat{Q}_H \) and gives priority to class 1 customers in the middle case so as to just achieve the contract. That is, as would be expected, the policy reduces the delay difference between the contract and non-contract customers as much as possible, while maintaining feasibility. We will refer to this limiting case as the Delay Difference-FCFS contract. Let \( Q^{DD-F} \) be the lower threshold \( \hat{Q}_M \) for the policy in regular traffic. Since all customers queue, \( Q^{DD-F} = \Lambda D \). A regular traffic policy that we hypothesize converges to the optimal policy for the Delay Difference-FCFS contract gives FCFS service if \( Q(t) \leq Q^{FCFS} \) or if \( Q(t) \geq Q_H \); if \( Q^{FCFS} < Q(t) < Q_H \), then class 1 is given priority if \( Q_1(t) \geq Q_L \). Rather than choosing an arbitrary \( c_1 \) for our numerical examples, we simulated this policy.

As the Delay Percentile contract is equivalent to the case of \( c_1 = 0 \), the results for cases with \( 0 < c_1 < \infty \) would lie between those of the Percentile Delay and the Delay Difference-FCFS contract.

In Table 2 we observe that the Delay Difference-FCFS contract reduces the contract customer delay by between 20% and 72% with an average over these cases of 42%. We do observe that in most of the cases there is a small increase in the CV of the delay, though in two cases it is reduced. However, the Convex Cost contract has lower delays and lower CVs as does the Combined Contract. The implication is that a contract that attempts to balance the delays experienced by both types of customers has a certain fairness to it and does limit the behavior of the firm under the Delay Percentile contract. However, contracts that attempt to minimize the exposure of the contract customers to extreme delays may be more in those customer’s interest than such Delay Difference contracts.

### 4 Discussion

In this paper we show that undesirable behavior from a call center contract customer’s perspective may in fact be rational for the call center, particularly under commonly used delay percentile contracts. We investigate the impact of this behavior on the customer and find that it can, in fact, be highly negative. In particular, the tails of delay, beyond the fixed point contracted for, may be very long. We also present alternate contract types focusing on convex functions of delay and the difference between delay for contract and non-contract customers. We show that contract customers can reduce the average delay and the variance
of delay they experience by focusing on these alternate contracts.

It is highly debatable whether in the real system that motivated this work the contract customers are even aware that they are not receiving “fair” service. If they choose to just evaluate their provider based on whether they meet the contract parameters, on the surface a reasonable metric, they may never know that in fact the tail of delay is very heavy. The heavy tails result from the rational behavior of the firm that assigns contract customers low priority when delays are largest. In order to address this problem, one recommendation we have for call center customers is that they evaluate their provider based on an explicit function of delay, such as average squared delay.

Alternatively, they may consider a contract that specifies several different percentiles of delay. That is, one might specify a contract with a lower percentile such as (80%, 20 sec.) and a higher percentile such as (95%, 1 minute). We repeated the analysis for the percentile delay contract, which added a second set of thresholds defining the service policy. For the numerical examples given, we simulated such a contract using the additional point of (95%, 1 min.) for cases 1-8 and (95%, 2 min.) for cases 9-12. We found that the expected waiting time for the contract customers decreased by 32% on average and the variance decreased by 27%, while the expected waiting time for the non-contract customers increased by 48% and the variance decreased by 11%. Thus such a contract reduces but does not eliminate the heavy-tails for the contract customers.

Further, we observe that the contract customers’ notion of “fairness” should not focus on assuring that all customers receive FCFS service except during busy periods, presumably when contract customers would be prioritized. Rather, we recommend that contract customers define fair treatment as having their contracted amount of delay observed as much as possible. That is, unfair treatment exists when the call center, upon realizing sufficient attention has already been given to the contract customers to meet the terms of the contract, gives lowest priority to the remaining contract customers. The Combined Contract we discuss, focusing on a convex function of excess delay, attempts to address such a notion of fairness.

Throughout this paper we assume Head-of-the-Line service. It is possible that even less desirable service (from a contract customer’s perspective) could be achieved by relaxing this assumption. In particular, it may be “optimal” to ignore any contract customers that have exceeded their contracted delay and serve them only when all other customers have left the system. However, such a policy seems rather extreme and perhaps impractical. A different type of analysis would be needed to study this type of policy because both the control theory proofs and state-space collapse arguments in this paper depend on the Head-of-the-Line assumption. In the single customer-type case such scheduling was examined in Koole (2005).

There are several extensions to this paper that might better clarify the use of call center contracts. We have not considered any model of contract competition in this paper. Clearly contract customers who are
not being treated well may choose to switch firms. However, as pointed out earlier, these customers may not even be aware of the long delay tail that they receive. A detailed model of contract customer behavior or a study of contracting company competition are interesting and challenging research problems that are left as the subject of future research.

We have implicitly assumed competition for non-contract customers because the call center wants to provide them with “good” service, modelled here as low mean delay. However, there is clearly room for an explicit model of competition. Some steps in this direction are provided by Armony and Haviv (2003) who consider delay competition between two service providers who determine service and pricing when there are different customer types. Of interest would be models of competition that allow customers to switch from being a non-contract to a contract customer.

Also of interest would be models that explicitly consider repeat customers, non-stationarity of arrival processes, abandonment from the queue, or different service times for the customer classes. Note that while there has been work on heavy-traffic limits under abandonment such limits are sometimes not impulse-exponential (see, e.g., Garnett et al. 2002). Also, the class of limits exhibiting state-space collapse becomes more difficult to characterize under Halfin-Whitt scalings when service times differ (see, e.g., Atar 2004). However, it is our intuition that none of these extensions will fundamentally change our findings that delay percentile contracts furnish incentives for undesirable firm behavior.

This paper has focused on practical implications of call center contracts rather than rigorous theory. The theoretical results are backed up by a simulation study that show the conclusions to be robust. Fully characterizing the set of limits that fit our definition of asymptotic optimality and rigorously proving state-space collapse under our regular-traffic policies is left as the subject of future research.

Acknowledgments

We are grateful to Mor Armony of New York University Stern School for assistance regarding prior literature on this topic. We would also like to acknowledge our original call center industry contact who was willing to share information on his company’s scheduling policies. Finally, we thank the two anonymous referees and the special issue editor, Ger Koole, whose comments improved the paper. This research was supported in part by a grant from NSERC and NSF grant DMI-0245382.
Appendix

Proof of Proposition 1: We can rewrite problem (2) in the following standard optimal control format:

\[
\begin{align*}
\min_{0 \leq u \leq 1} J &= \int_0^\infty x(1 - u(x))dF_Q(x) \\
\text{s.t.} &
\begin{align*}
\dot{v}(x) &= I_{xu(x) \leq \hat{Q}_L} \alpha \beta e^{-\beta x} \\
v(0) &= 1 - \alpha \\
v(\infty) &= \gamma
\end{align*}
\end{align*}
\]

From the structure of the problem we observe that if \(F^c_Q(\hat{Q}_L) \geq \gamma\), \(u^*(x) = 1\) for all \(x\). Otherwise, the objective function and the form of the indicator function imply \(u^*(x) = 1\) for \(x \leq \hat{Q}_L\), the constraint \(v(\infty) \geq \gamma\) is tight in any optimal solution and considering the indicator function and that the objective is linear in \(u(x)\), the solution is bang-bang with \(u(x) = 1\) or \(u(x) = \hat{Q}_L/x\). Letting \(y(x) = I_{xu(x) \leq \hat{Q}_L}\), \(\delta = F^c_Q(\hat{Q}_L)\) and \(\eta = \gamma - \delta\), we can rewrite the problem as

\[
\begin{align*}
\min_{y(x) \in \{0, 1\}} J &= -\hat{Q}_L \eta + \int_{\hat{Q}_L}^\infty x y(x) dF_Q(x) \\
\text{s.t.} &
\begin{align*}
\dot{v}(x) &= y(x)f_Q(x) \\
v(\hat{Q}_L) &= \delta \\
v(\infty) &= \gamma
\end{align*}
\end{align*}
\]

Consider the linear relaxation of the problem (0 \(\leq y(x) \leq 1\)) for which the Hamiltonian is

\[
H(v, u, x, \nu) = -\hat{Q}_L \eta + (\nu(x) - x)y(x)f_Q(x)
\]

with \(\dot{\nu}(x) = H_v = 0\) so \(\nu(x) \equiv \nu\) is a constant. As the Hamiltonian is linear in \(y(x)\), we maximize it by letting \(y(x) = 1\) if \(\nu \geq x\) and \(y(x) = 0\) otherwise. Therefore, letting \(\nu = \hat{Q}_H\) we have \(\int_{\hat{Q}_L}^{\hat{Q}_H} dF_Q x = \eta\) or \(F^c_Q(\hat{Q}_H) = \gamma\).

Proof of Observation 1: Comparing the approximate expected queue lengths for each class of customer and using Little’s Law provides the result.

Proof of Observation 2: Considering problem (5), as in the case of the proof of Proposition 1, we can reason that the constraint is tight and the solution bang-bang with \(u(x) = 1\) or \(u(x) = \hat{Q}_L/x\). Letting \(y(x) = 1\) if \(u(x) = 1\) and 0 otherwise, and \(\nu(x)\), be the Lagrangian multiplier on the constraint, we can repeat the proof of Proposition 1, to find the Hamiltonian of the associated control problem as

\[
H(v, u, x, \nu) = -cy(x) + (x - Q_L)(1 - y(x))\nu,
\]

25
which gives the result that \( u(x) = 1 \) for \( x \leq \hat{Q}_L \), \( u(x) = \hat{Q}_L/x \) for \( \hat{Q}_L \leq x < \hat{Q}_K \), and \( u(x) = 1 \) for \( x \geq \hat{Q}_K \), where \( \hat{Q}_K \) is determined by the constraint. Note that since the Hamiltonian is equivalent to placing a linear cost on the delay for non-contract customers, the formulation minimizing \( c_1E[1_{W_1 \geq D}] + c_2E[W_2] \) is also equivalent.

**Proof of Observation 3:** Noting the constraint to be tight in any optimal solution, \( E[\hat{Q}xu(x)] = E[\hat{Q}] - K' \).

**Proof of Proposition 2:** The corresponding Hamiltonian is

\[
H = (\phi(xu(x)) + \nu(x(1 - u(x)))) f_{\hat{Q}}(x),
\]

which, considering \( \partial H/\partial u \), implies \( u(x) = 1 \) if \( \phi'(x) < \nu \) and \( u(x) = 0 \) otherwise. Letting \( \hat{Q}^* \) solve \( \int_{\hat{Q}}^{\hat{Q}^*} x dF_{\hat{Q}}(x) = K' \) implies \( \nu = \phi'(\hat{Q}^*) \). Thus, the firm would give priority to non-contract customers when \( \hat{Q}(t) < \hat{Q}^* \) and priority to contract customers otherwise.

**Proof of Observation 4:** From the optimal policy for each contract, we know

\[
\int_{\hat{Q}_H}^{\infty} f_{\hat{Q}}(x)dx = \int_{\hat{Q}_L}^{\hat{Q}^*} f_{\hat{Q}}(x)dx = 1 - \gamma
\]

Then we find

\[
\frac{1}{\alpha} \left( E[\hat{Q}^{PP}_1] - E[\hat{Q}^{CC}_1] \right) = \int_{\hat{Q}_L}^{\hat{Q}_H} Q_Lf_{\hat{Q}}(x)dx + \int_{\hat{Q}_H}^{\infty} x f_{\hat{Q}}(x)dx - \int_{\hat{Q}_L}^{\hat{Q}^*} x f_{\hat{Q}}(x)dx
\]

\[
= \hat{Q}_L(e^{-\beta Q_H} - e^{-\beta Q_H}) + \hat{Q}_H e^{-\beta Q_H} + \int_{\hat{Q}_H}^{\infty} x f_{\hat{Q}}(x)dx
\]

\[
- \hat{Q}_L e^{-\beta Q_L} + \hat{Q}^* e^{-\beta Q^*} - \int_{\hat{Q}_L}^{\hat{Q}^*} x f_{\hat{Q}}(x)dx
\]

\[
= (\hat{Q}_H - \hat{Q}_L)e^{-\beta Q_H} + \hat{Q}^* e^{-\beta Q^*}
\]

\[
> 0
\]

**Proof of Proposition 3** From Proposition 2, \( u(x) = 1 \) for \( x < \hat{Q}' \) and \( u(x) = 0 \) otherwise, for some \( \hat{Q}' \). However, as the class 1 cost is convex in \( (\hat{Q} - \hat{Q}_L)^+ \), letting \( \hat{Q}_1 = \hat{Q}_L \) for \( \hat{Q} > \hat{Q}' \) has no cost with respect to class 1 and the linear term in \( E[\hat{Q}_2] \) is minimized by doing so. Therefore \( u(x) = \hat{Q}_L/x \) for \( x > \hat{Q}' \). Placing the constraint \( P(\hat{Q}_1 \leq \hat{Q}_L) \geq \gamma \) fixes \( \hat{Q}' \) at \( \hat{Q}^* \) and is the only feasible point in the set of unconstrained optimal policies and so is optimal.
Proof of Observation 5: That the first inequality holds is obvious from construction. From the proof of Observation 4, we find

\[ E[Q_1^B] - E[Q_1^D] = (Q_H - Q_L)e^{-\beta_2\hat{Q}^*} + (Q^* - \hat{Q})e^{-\beta_1\hat{Q}^*} \]

where the last term in the first line reflects that policy \( u(x) = \hat{Q}_L/x \) for \( x > \hat{Q}^* \).

Proof of Proposition 4: Let \( J(u) = c_1 \left( \frac{x(u)}{\lambda_1} - \frac{x(1-u(x))}{\lambda_2} \right)^2 + c_2 x (1-u(x)) \). For the unrestricted problem \( \min_{0 \leq u(x) \leq 1} J(u) \), \( dJ/du = 0 \) implies the solution \( u = a/x + p_1^* \), noting \( d^2J/du^2 = 2c_1(x/\lambda_1 + x/\lambda_2)^2 \). Observe that \( x(a/x + p_1^*) \leq \hat{Q}_L \) on \( x \leq \hat{Q}_M \). For \( x > \hat{Q}_M \), we observe the constraint implies the bang-bang solution \( u(x) = \hat{Q}_L/x \) or \( u(x) = a/x + p_1^* \). Observe

\[ J \left( \frac{\hat{Q}_L}{x} \right) = \frac{c_1}{(\lambda_2)^2} x^2 + \frac{c_2 (\lambda_1^2 \lambda_2^2 - 2c_1 \lambda_1^2 \lambda_2) \hat{Q}_L}{\lambda_1 + \lambda_2} \left( \frac{c_1 (\lambda_1^2 \lambda_2^2 - 2c_1 \lambda_1^2 \lambda_2) \hat{Q}_L}{\lambda_1^2 \lambda_2} \right) \]

is quadratic in \( x \) and

\[ J(a/x + p_1^*) = \frac{c_2 x}{\lambda_1 + \lambda_2} - \frac{c_2^2 (\lambda_1)^2}{4c_1 (\lambda_1 + \lambda_2)^2} \]

is linear in \( x \). Further \( J(\hat{Q}_L/x) \geq J(a/x + p_1^*) \) for \( x \geq \hat{Q}_M \), equality holding only at \( \hat{Q}_M \). Noting the optimal solution has \( \frac{x(u(x))}{\lambda_1} \geq \frac{x(1-u(x))}{\lambda_2} \) when \( u(x) = a/x + p_1^* \), and is independent of \( c_1 \) on those intervals, the solution also minimizes \( J'(u) = c_1 \left( \frac{x(u(x))}{\lambda_1} - \frac{x(1-u(x))}{\lambda_2} \right)^2 + c_2 x (1-u(x)) \).

Since \( \hat{Q}^* \) is an impulse-exponential random variable, the constraint \( E_{\hat{Q}} \left[ 1_{x(u(x)) \leq \hat{Q}_L} \right] \geq \gamma \) implies there exists \( \hat{Q}' \) for which \( J(u) \) is minimized by letting \( u(x) = \hat{Q}_L/x \) for \( x \leq \hat{Q}' \) and \( u(x) = a/x + p_1^* \) for \( x > \hat{Q}' \).

But we require \( \int_0^{\hat{Q}'} dF_{\hat{Q}}(x) = \gamma \) so as in Proposition 1, \( \hat{Q}' = \hat{Q}_H \).
References


Fluid Model Convergence

Instead of formally proving state-space collapse (which would also require formally setting up the sequence of systems and corresponding scalings) we instead show that a fluid system operated under $\pi_A$ converges to a fixed point in finite time, where each queue has the quantity of work predicted by state-space collapse. As such a proof forms the core of most of the prior state-space collapse results cited in Sections 2.1 and 2.2, this provides strong evidence for the hypothesized convergence.

We consider a fluid version of the original problem where the queueing network is modeled by a piecewise-linear fluid system where workload arrives and is depleted in a deterministic and continuous manner in each queue. In other words, workload is represented as a fluid that flows continuously through the system. Such fluid sample paths are often derived as an appropriate (law of large numbers) scaled limits of the original system (see, e.g., Section 8 of Mandelbaum and Stolyar 2004). Because we are not formally proving state-space collapse, we instead present the fluid model without the formal limits.

We let $r_i(t)$ be the rate applied by the server to class $i$ at time $t$, $i = 1, 2$, $t \geq 0$. Assume that capacity is $\bar{r}$, so that $\sum_{i=1}^{2} r_i(t) \leq \bar{r}$. Under conventional heavy-traffic limit theorems the appropriate value for $\bar{r}$ would be $N\bar{\mu}$; here, we merely assume some finite value. We assume that fluid flows into queue $i$ at rate $p_i\bar{r}$, $i = 1, 2$, so that the system is operating in heavy-traffic. Let $\bar{X}(t)$ be the work in the system at time $t$ so that if $\sum_{i=1}^{2} r_i(t) = \bar{r}$ for all $t \geq 0$ then $\bar{X}(t) = \bar{X}(0)$ for all $t \geq 0$.

The fluid version of the two-threshold policy $\pi_A$ has thresholds $X_L$ and $X_H$ (which are appropriately scaled versions of $Q_L$ and $Q_H$, respectively). Under this policy $r_2(t) = \bar{r}$ (and hence $r_1(t) = 0$) if $\bar{X}_2(t) > 0$ and $\bar{X}(t) \leq X_L$ or $\bar{X}(t) \geq X_H$. If $\bar{X}_2(t) = 0$ and $\bar{X}(t) \leq X_L$ or $\bar{X}(t) \geq X_H$ then $r_2(t) = \bar{p}_2\bar{r}$ and $r_1(t) = \bar{p}_1\bar{r}$.

For $X_L < \bar{X}(0) < X_H$, if $\bar{X}_1(0) \geq X_L$ then $r_1(t) = \bar{r}$ otherwise $r_1(t) = 0$.

Proposition 5 Define $\bar{T} = \max(\bar{X}(0)/(\bar{p}_1\bar{r}), (\bar{X}(0) - X_L)/(\bar{p}_2\bar{r}))$, then for $t \geq \bar{T}$

$$
\bar{X}_1(t) = \begin{cases} 
\bar{X}(0) & \text{if } \bar{X}(0) \leq X_L \\
X_L & \text{if } X_L < \bar{X}(0) < X_H \\
\bar{X}(0) & \text{if } \bar{X}(0) \geq X_H 
\end{cases}
$$

As described earlier, many authors have used the result that, if the fluid scale version of the model converges in finite time then state-space collapse occurs in the diffusion limit (to the queue allocations described by the fixed point of the fluid limit). We have therefore provided strong, but not conclusive, evidence that the policy $\pi_A$ is asymptotically optimal among all policies that yield heavy-traffic limits of the type outlined in Section 2.
To make this conclusion rigorous we would need to formally set up the limit theorems. However, in the theorem the form of the fixed point depends on initial workload. This is consistent with Armony and Maglaras (2004a, 2004b) but not with Bramson (1998). However, unlike Armony and Maglaras (2004a, 2004b), there is a discontinuity in the fluid process as total workload crosses a $X_H$. It is not clear whether their proof technique would carry over to this discontinuous setting. This is left as the subject of future research.