OPERATIONS RESEARCH
Vol. 52, No. 4, July–August 2004, pp. 527–545
ISSN 0030-364X | EISSN 1526-5463 | 04 | 5204 | 0527

DOI 10.1287/opre.1040.0123

Contact Centers with a Call-Back Option and Real-Time Delay Information

Mor Armony
Stern School of Business, New York University, 44 West 4th Street, New York, New York 10012, marmony@stern.nyu.edu

Constantinos Maglaras
Columbia Business School, 409 Uris Hall, 3022 Broadway, New York, New York 10027, c.maglaras@columbia.edu

Motivated by practices in customer contact centers, we consider a system that offers two modes of service: real-time and postponed with a delay guarantee. Customers are informed of anticipated delays and select their preferred option of service. The resulting system is a multiclass, multiserver queueing system with state-dependent arrival rates. We propose an estimation scheme for the anticipated real-time delay that is asymptotically correct, and a routing policy that is asymptotically optimal in the sense that it minimizes real-time delay subject to the deadline of the postponed service mode. We also show that our proposed state-dependent scheme performs better than a system in which customers make decisions based on steady-state waiting-time information. Our results are derived using an asymptotic analysis based on "many-server" limits for systems with state-dependent parameters.

Subject classifications: service networks; service level guarantees; multiclass queueing systems; call-back option; call centers; Halfin-Whitt regime; real-time delay notification.
Area of review: Manufacturing, Service, and Supply Chain Operations.
History: Received June 2002; revision received January 2003; accepted July 2003.

1. Introduction

Many organizations use customer contact centers as an important channel of communication with their customers. Such centers have limited resources and face highly unpredictable demand that often result in long waits for their customers. To improve the customer service levels and alleviate congestion, contact centers have recently started experimenting by: (a) informing arriving customers (callers) about anticipated delays, and (b) adding alternate service modes. The most common example of such a system is a telephone call center that, in addition to regular service, also offers a call-back option, whereby customers may register a request and the system will call them back within a prespecified amount of time. In such systems, callers use the announced information and their personal preferences to decide whether to wait for real-time service, leave a service request, or simply balk.

While such initiatives are not surprising, they do raise some theoretical and practical questions. Specifically, how can the system accurately estimate the anticipated delay for real-time service? What is the routing rule that optimizes system behavior subject to the quality-of-service guarantee offered to the "call-back" customers? In terms of performance, one would expect that these initiatives will lead to some form of *load balancing* by shifting service requests from one channel to the other when the system is congested. How much benefit (if any) can be obtained by announcing this state-dependent information to the customers? What is the value of the call-back option?

We make some initial progress in addressing these questions by focusing on a model for a contact center with two service modes: real-time and postponed with a delay guarantee. We propose a scheme that allows the call center to provide accurate delay estimates and an accompanying routing rule that guarantees that the postponed service is offered within the prespecified deadline. We justify these choices through an asymptotic analysis that also provides an approximation for the closed-loop behavior of the system. Finally, we compare this system's performance with that of a system that announces the steady-state expected waiting-time information to the arriving customers. In this context, we show that the use of state-dependent information increases the overall system utilization while providing better quality of service to the real-time customers, and maintaining the same level of service for the call-back customers.

The system under consideration can be modeled as a two-class $M/M/N$ queueing system, where customers are informed upon arrival of their (state-dependent) waiting time for real-time service (class 1) and the delay bound for the call-back service (class 2), and make decisions by maximizing the relative utility associated with these two options or with balking. Because customers' decisions are based on state-dependent information, the resulting arrival rates to each class are also state dependent. This leads to a major challenge in estimating the anticipated delay for real-time service. Although such an estimation is trivial in a single-class system operating under a FIFO policy, it becomes quite complex in multiclass systems where,

depending on the routing policy, future arrivals may affect the waiting time of customers already in queue. This is also the case in our system, where the delay constraint for call-back customers entangles the operation of the two service modes, and the anticipated delay for a class 1 customer may indeed depend on future arrivals to class 2. Delay estimation in systems where the waiting or service time depends on future arrivals tends to be a difficult task (see, for example, Whitt 1999b, Ward and Whitt 2000, Hassin and Haviv 1997, Bitran and Caldentey 2002).

In addition, even with an accurate estimate of the anticipated delay for real-time service, the performance analysis of a two-dimensional birth-death model with state-dependent transition rates is both analytically and numerically complex. Instead, the analysis in this paper focuses on the so-called many-server, heavy-traffic regime, first studied by Halfin and Whitt (1981). This extends previous work by the authors (Armony and Maglaras 2004) that analyzed customer contact centers with a call-back option but where customers make their service choice based on steady-state delay information, which they gain from experience or from published delay performance. The key contributions of this paper are the following:

(1) In terms of solution methodology, we first show that if the system is not significantly over- or under-capacitated, then rational customer choice behavior based on anticipated delay information will place the system in heavy traffic. Second, we propose a delay estimation rule that is *asymptotically consistent*, and an asymptotically optimal routing rule that guarantees the delay constraint for the call-back customers. The former implies that the actual delay experienced by class 1 customers asymptotically agrees with what was announced to them upon their arrival. The proposed estimator is simple to implement and is derived based on insights extracted from the asymptotic behavior of multiserver systems in the Halfin-Whitt regime.

The criterion of an *asymptotically consistent* estimator is interesting in its own right, and may have broader applicability, for example, in multiclass queueing systems with leadtime quotation that leads to state-dependent arrival streams.

(2) In terms of analysis, we use a nontrivial application of a classical result on diffusion approximations, Stone's criterion, to establish a limit theorem for the one-dimensional total queue-length process, and then derive the class level queue length and waiting time behavior by establishing the appropriate state-space collapse property.

(3) These analytical results lead to closed form performance approximations that, in turn, yield several insights about the operation and design of such systems. (i) The state of the system evolves in a much slower time scale than the waiting times experienced by the customers. This implies that the state of the system stays constant within the time that callers spend in queue, making the estimation of their anticipated delay relatively simple and robust. (ii) Announcing state-dependent information improves performance. In particular, it increases overall resource utilization while simultaneously improving the quality of service experienced by customers selecting the real-time service mode. A similar insight was derived by Whitt (1999a) for a single-class model of a call center. (iii) The closed form performance approximations allow one to study the value of the call-back option, and lead to simple numerical recipes for optimizing system behavior by introducing a call-back channel with a suitable delay guarantee. (iv) The commonly used "square-root" staffing rule is shown to apply in our system with two service channels and delay guarantees, and a closed-form characterization of the appropriate staffing level is provided.

The remainder of this paper is structured as follows. We conclude this section with a short literature survey. Section 2 describes the basic model. Section 3 provides some background results on asymptotic analysis of systems with stationary (state independent) arrival rates. Section 4 analyzes the system of interest, and §5 compares its performance to one that announces steady-state delay information. Section 6 provides some concluding remarks.

The literature related to our work spans three main areas. The first is concerned with the analysis of multiserver systems motivated by call center applications. The majority of this work focuses on single-class systems. For example, see Kolesar and Green (1998) and the references therein. The papers by Brandt and Brandt (1999) and Gans and Zhou (2003) analyzed two-class models with static parameters, but with routing issues that are close to ours. The first paper analyzed a given policy using Markov chain methods, while the second used an MDP formulation to characterize the optimal routing rule.

The second area of literature that is close to our work is related to the asymptotic analysis of multiserver systems pioneered by Halfin and Whitt (1981). Recent work along these lines include Jennings et al. (1996), Garnett et al. (2002), Puhalskii and Reiman (2002), Borst et al. (2004), Armony and Maglaras (2004), and Whitt (2003). The diffusion analysis with state-dependent parameters uses Stone's criterion (1963) (see, for example, Iglehart 1965), related to which is Mandelbaum and Pats (1995).

The third body of literature is on leadtime quotation in production systems that is related to the problem of announcing waiting times. Close references are Duenyas and Hopp (1995), Duenyas (1995), Plambeck (2001), and Dobson and Pinker (2002). In some of these papers the issue of "optimal" leadtime quotation is considered, where the benefits of both overestimating and underestimating the leadtimes are explicitly investigated. Although we recognize the importance of this trade-off, we focus in this paper on quoting waiting times as accurately as possible, leaving the question of "optimal" quotation to future research. As it will turn out, our estimate will be asymptotically exact.

Finally, we comment briefly on the results obtained by Whitt (1999a) that studied the effect of announcing state-dependent delay information in a single-class, multiserver

system by comparing it to a model where customers first join the system but later abandon it if their service does not start within a specified time (that is customer specific). In contrast, our model has two service classes, customers are assumed to make joining decisions based on either state-dependent or steady-state delay information, and there is no abandonment. The system that announces steady-state information, which is used as a benchmark, is a stochastic equilibrium model that incorporates the customers' reaction to delay.

2. The Model

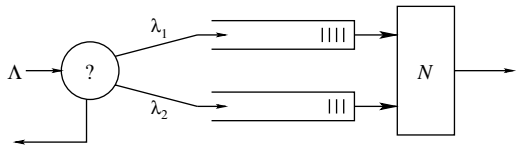
The service system (depicted in Figure 1) has N statistically identical servers. It provides two types of service: (a) *real-time* service, where users join a first-in-first-out (FIFO) queue (queue 1, here), and (b) *postponed* (call-back) service, where users leave a message and the system calls them back within D_2 time units (this is queue 2). The upper bound on the waiting time for class 2 service requests is not meaningful in a conventional sense, because the latter are unbounded random variables. However, as we will show, this constraint can be guaranteed in an appropriate asymptotic regime as the number of servers increases and system utilization goes to 1. Both classes have identical processing requirements, and service times are independent, exponential random variables with mean m (and service rate $\mu = 1/m$). The system parameters N and D_2 are assumed to be fixed.

We denote by $Q_i(t)$ and $Z_i(t)$ the number of class i customers waiting in queue and in the system at time t, respectively. Let $s(t) = (Q_1(t), Q_2(t), Z_1(t) + Z_2(t))$ be the state of the system at time t, and set $S(t) = \{s(\tau)\colon \tau \leqslant t\}$ be the history of the process up to time t. (Note that (Q_1, Q_2) are insufficient as state descriptors because when $Q_1 + Q_2 = 0$ the transition probabilities depend on the value of $Z_1 + Z_2$.)

Customer Characteristics

Customers arrive according to a Poisson process with rate Λ. Upon arrival they have three choices: (i) join queue 1 and wait to be processed, (ii) leave a message for postponed service in queue 2, or (iii) balk and do not join the system. (We assume that customers who decide to balk do not retry later, or, if they do, the corresponding retrial rate is negligible.) Arriving customers are informed of (a) the state-dependent anticipated waiting time in class 1, $w_1(S)$, and (b) the delay D_2 within which they will receive a call-back should they select option (ii). Based on this information they decide whether to join the system and what type of service to request. The key trade-off they face between real-time and postponed-but-guaranteed service is analogous to the trade-off between "best-effort" and "guaranteed" type of service in communication networks. It is related to the cost of waiting that each customer associates with each of these service modes. We denote by $\lambda_1(S), \lambda_2(S)$, and $\lambda_0(S)$ the *state-dependent* rates at which customers join class 1, class 2, or balk, respectively. Clearly, $\Lambda = \lambda_1(S) + \lambda_2(S) + \lambda_0(S)$.

Figure 1. A system with two service channels.



Customer Choice Model

We assume that there is a continuum of customer types, indexed by τ, that are differentiated by their preferences. User preferences are determined as follows:

(a) The utility for real-time service with anticipated delay w_1 is $u_1(w_1, \tau)$.

(b) The utility for leaving a request for a call-back within D_2 time units is $u_2(D_2, \tau)$.

We assume that $u_1(w_1, \tau)$ and $u_2(D_2, \tau)$ are nonincreasing with respect to the first variables, continuously differentiable with respect to both variables, and $u_i(\infty, \tau) < 0$, i.e., the utility is negative if w_1 or D_2 are sufficiently large, because in such cases the cost of waiting exceeds the value obtained by receiving service, making such choices undesirable. For $i = 1, 2$, $u_i(0, \tau)$ represents the utility for receiving immediate service (no wait), which depreciates as the customer has to wait either on-line ($i = 1$) or off-line to be called back ($i = 2$). Without loss of generality, we assume that the utility of not joining is zero; that is, $u_0 = 0$. Customers choose the type of service that maximizes their own utility according to $\max\{0, u_1(w_1, \tau), u_2(D_2, \tau)\}$; that is, a type τ customer will join queue 1 if $u_1(w_1, \tau) \geqslant u_2(D_2, \tau)$ and $u_1(w_1, \tau) \geqslant 0$, leave a service request in queue 2 if $u_2(D_2, \tau) > u_1(w_1, \tau)$ and $u_2(D_2, \tau) \geqslant 0$, and balk if $u_1(w_1, \tau), u_2(D_2, \tau) < 0$.

Finally, the customer type is a random variable. Let P^τ be the probability distribution over the set of customer types. We assume that the type distribution has a continuous density function on its support and that for all finite $x \geqslant 0$, $\mathbf{P}^\tau(u_i(x, \tau) \geqslant 0) > 0$; for any set A, $\mathbf{P}^\tau(A)$ denotes the probability of that set under the distribution P^τ. The type of each customer is chosen according to P^τ and is independent of all other customer types. Given this setup,

$$\lambda_1(S) \triangleq \lambda_1(w_1(S), D_2) = \Lambda \mathbf{P}^\tau(u_1(w_1(S), \tau) \geqslant u_2(D_2, \tau) \quad \text{and} \quad u_1(w_1(S), \tau) \geqslant 0), \tag{1}$$

$$\lambda_2(S) \triangleq \lambda_2(w_1(S), D_2) = \Lambda \mathbf{P}^\tau(u_2(D_2, \tau) > u_1(w_1(S), \tau) \quad \text{and} \quad u_2(D_2, \tau) \geqslant 0), \tag{2}$$

and $\lambda_0(S) \triangleq \lambda_0(w_1(S), D_2) = \Lambda - \lambda_1(w_1(S), D_2) - \lambda_2(w_1(S), D_2)$. The standing assumptions on τ, $\mathbf{P}^\tau$ and $u_i(\cdot, \tau)$ imply that $\lambda_i(\cdot, \cdot)$ is continuously differentiable with respect to both arguments. Also, we define $\Lambda_{\text{eff}} := \lambda_1(0, 0) + \lambda_2(0, 0)$ to be the maximum rate at which

customers may choose to join the system, which is achieved when $w_1 = 0$ and $D_2 = 0$. In general, $\Lambda_{\text{eff}} \leqslant \Lambda$ represents the "effective" load for the system. We will assume that $\mathbf{P}^\tau(u_1(0,\tau) > u_2(0,\tau)) > 0$ and that $\mathbf{P}^\tau(u_2(0,\tau) > u_1(0,\tau)) > 0$. The interpretation for this condition is that the two service modes are not perfect substitutes of each other, and this is reflected in their respective utilities.

A simple example, which is used later on for illustrative purposes, is one with linear waiting costs and the specific structure of the Multinomial Logit Model (MNL) (see Anderson et al. 1996, §2.6). In this case, for appropriate constants r_i and c_i, $i = 1, 2$, we have

$$\bar{u}_1(S) = r_1 - c_1 w_1(S) \quad \text{and} \quad \bar{u}_2(S) = r_2 - c_2 D_2,$$

and the total utility for each choice is given by $u_i(S,\tau) = \bar{u}_i(S) + \tau_i$, where τ_i are IID double exponential (Gumbel) distributed with parameter ν; that is, $\tau = [\tau_1, \tau_2]$ defines the random portion of the two utility functions that differentiates between customer types. With respect to the parameters of this model, it is natural (but not necessary) to assume that $r_1 \geqslant r_2$ and $c_1 > c_2$; that is, customers mind more when they are waiting in queue 1, rather than waiting off-line for their request to be processed. For the MNL choice model, (1)–(2) simplify to

$$\lambda_i(S) = \Lambda \frac{e^{\bar{u}_i(S)/\nu}}{1 + \sum_{j=1,2} e^{\bar{u}_j(S)/\nu}}. \tag{3}$$

To complete the model description we need to specify how the system manager estimates the anticipated delay for real-time service, and what is the routing rule that determines which type of customer to serve next whenever a server becomes available; the two are obviously related.

Routing Policy

A natural candidate to consider would give class 2 calls service priority whenever one of these jobs is about to violate its delay deadline, and the rest of the time give priority to class 1. The obvious shortcoming of this policy is the complexity of the state descriptor, which must keep track of the age of all class 2 jobs in the system, that makes performance analysis of the system hard. In this paper, we propose a policy that only uses queue length information, and gives priority to class 2 when its queue length exceeds a certain threshold, and to class 1, otherwise. As we argue in the sequel, the queue length will act as an effective and accurate proxy for the age of the jobs in queue, thus imitating the age-based policy described above.

Let $A_i(t)$ be the number of customers that have arrived into queue i in $[0, t]$. Using an observation made in Maglaras and Van Mieghem (2004), we have that

$$W_2(t) \leqslant D_2 \quad \forall t \quad \Leftrightarrow \quad Q_2(t) \leqslant A_2(t) - A_2(t - D_2) \quad \forall t,$$

i.e., no class 2 customer has been waiting for more than D_2 if and only if all the customers currently in queue 2 arrived within the last D_2 time units. Hence, the appropriate threshold to use is

$$\theta(t) = A_2(t) - A_2(t - D_2),$$

and the corresponding policy is specified as follows:

If $Q_2(t) \geqslant \theta(t)$, give priority to class 2,
else give priority to class 1.

Note that it is easy to implement policies that keep track of $\theta(t)$; this is particularly true for call centers, where such information is readily available. This policy cannot guarantee that the delay constraint will be satisfied for *all* class 2 customers, nor can any alternative policy. This is due to the stochasticity and unboundedness of the waiting times. However, as will be shown later, this policy does guarantee that *asymptotically* as the size of the system grows large, the delay guarantee will indeed hold for all class 2 customers, with probability 1.

Estimation of Anticipated Delay

The task of estimating the expected delay for real-time service conditional on the current state of the system is quite involved. Indeed, the dependence of the arrival rates on the state of the system and the structure of the proposed routing policy make this calculation very complex because current class 1 delays depend on future class 2 arrivals, and both are functions of the state that is changing with time. However, if one focuses on large, heavily loaded systems, which typify customer contact centers, the situation simplifies dramatically. This is due to the following key observation which will be explained in detail later on: Large multiserver systems enjoy a form of *statistical economies of scale*; in particular, the waiting times of customers in the real-time queue decrease to zero, even if the system is approaching heavy traffic. Moreover, the state of the system (number of busy servers, and number of customers in queue) does not change significantly during each such short waiting period.

So, assuming optimistically that the state and arrival rates are indeed constant over the time a customer spends in queue 1, one can compute the state-dependent anticipated delay as follows. Given the class 1 queue length, q_1, and the arrival rate, λ_1, a local version of Little's law (to be justified asymptotically later on) shows that the class 1 waiting time may be approximated by

$$\hat{w}_1(q) = \frac{q_1}{\lambda_1}.$$

We conclude with some comments about our model. First, we assume that customers cannot abandon or jockey once they join the system. This is reasonable provided that the estimation of their anticipated delay is fairly accurate, which will turn out to be the case. Second, the total arrival

rate Λ is assumed to be time invariant. This is not realistic, because in most service systems, such as call centers, there is a very pronounced variation depending on time of day, day of week, promotional offers, etc.; see Green and Kolesar (1991). However, the waiting time estimation and the asymptotic analysis of the next few sections extend very naturally to the nonstationary setting, because both settings do not require a steady-state analysis. Third, the assumption that the service time requirements of the two job classes are equal is clearly a modeling idealization, and is imposed for analytic tractability purposes. (The analytical complexity of multiclass, multiserver systems when the service rates are different is well known.) While all the structural insights extracted in this paper apply for the case with different μs, some of the particular details of the approximating distributions will change; some recent results by Maglaras and Zeevi (2002) outline how this is done.

Finally, one could consider a system that announces different types of information for the two service options, e.g., conditioned on the state, announce the 80th percentile of the waiting time distribution for class 1 service rather than its expected value, and assume that the customers have an appropriate utility function to process this type of information. Such alternatives will not be considered for the reason we explain below. As will be shown in §4, the proposed estimate that the system announces is asymptotically exact for each customer that goes through the system (this is pointwise equality and not just in expectation). In our asymptotic regime, the waiting time distribution degenerates to this point estimate (this is due to the separation of time scales briefly mentioned earlier), which, in turn, provides an appropriate approximation for large capacity systems. Specifically, in large contact centers, even if the system was announcing different types of information regarding the waiting time distribution for class 1, our asymptotic analysis would show that the effect of such different signals on system performance and customer behavior is negligible. Moreover, the resulting asymptotic approximations would end up being the same as under $\widehat{w}_1(q)$.

3. Background: Large Capacity Asymptotics for Multiserver Systems

Typical contact centers are large in size and tend to be heavily loaded. Our approach uses an asymptotic analysis motivated by these two observations. This section provides some background about the asymptotic behavior of multiserver systems as N grows large and the traffic intensity approaches one. Both will prove useful in §4.

Operating Regimes

The "physical" modes of operation for the a multiserver system as N grows large can be classified by focusing on the probability that a randomly selected customer arriving to the system will have to wait before getting served. Following Gans et al. (2003), we consider three modes of operation:

- **Efficiency-Driven Regime:** the system is under-capacitated and customers almost always wait, $\mathbf{P}(\text{wait} > 0) \approx 1$.
- **Quality and Efficiency-Driven (QED) Regime:** the system's capacity is *balanced* and customers may have to wait but not always, $\mathbf{P}(\text{wait} > 0) \approx \alpha \in (0, 1)$; we also refer to this as the **Halfin-Whitt regime**.
- **Quality-Driven Regime:** the system is over-capacitated and customers almost never wait, $\mathbf{P}(\text{wait} > 0) \approx 0$.

The efficiency-driven regime under-emphasizes congestion effects, the quality-driven regime focuses on service quality, while the QED regime achieves a balance between operating costs and quality of service. As advocated by Garnett et al. (2002) this seems to be the natural operating regime to consider. Furthermore, Borst et al. (2004) and Maglaras and Zeevi (2003) have shown that this is the economically optimal regime in single-class multiserver models where the system manager optimally selects the capacity level and/or the price users must pay to gain access to the system.

The remainder of this section provides some additional background on the QED or Halfin-Whitt regime, under the simplifying assumption that the arrival rates are stationary and exogenously given, i.e., do not depend on W_1, D_2, or the state of the system in any way. This summary suggests natural scaling relationships for the model studied in this paper, which will be exploited later on.

The Halfin-Whitt Regime

Consider a system with N servers and denote by $Z_i^N(t)$ and $Q_i^N(t)$ the total number of class i jobs present in the system or in queue at time t, respectively; a superscript N will be attached to all relevant quantities to denote their dependence on the size of the system. The class level arrival rates are λ_1^N and λ_2^N, and the aggregate arrival rate is given by $\lambda_a^N = \lambda_1^N + \lambda_2^N$. The first observation that we make is that the total number of customers in the system, given by $Z^N(t) = Z_1^N(t) + Z_2^N(t)$, behaves precisely like an $M/M/N$ system with arrival rate λ_a^N. In particular, the total number of customers in the system is independent of the routing rule, provided that it is nonidling. The specific routing decisions will affect the class-level queue-length processes.

Following the informal discussion given above, let us define the probability of congestion as $\mathbf{P}(\text{wait} > 0) = \mathbf{P}(Z^N \geqslant N)$, where the notation Z_i^N and Q_i^N without a time argument denote these random variables in steady-state. In their 1981 paper (Proposition 1), Halfin and Whitt established that

$$\lim_{N\to\infty} \mathbf{P}(\text{wait} > 0) = \alpha \in (0,1) \quad \text{iff}$$

$$\rho^N := \frac{\lambda_a^N}{N\mu} = 1 - \frac{\beta}{\sqrt{N}} + o\left(\frac{1}{\sqrt{N}}\right), \quad \beta > 0 \tag{4}$$

(i.e., $\sqrt{N}(1-\rho^N) \to \beta$ as $N \to \infty$), where $\alpha = [1 + \sqrt{2\pi}\beta\Phi(\beta)e^{\beta^2/2}]^{-1}$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. This condition provides a

precise articulation of the asymptotic parameter setting that will give rise to the Halfin-Whitt regime. Assume that (4) holds and define

$$X^N(t) = \frac{(Z_1^N(t) + Z_2^N(t)) - N}{\sqrt{N}}.$$

Halfin and Whitt established that $X^N \Rightarrow \widetilde{X}$, where $\widetilde{X}$ is a well-defined diffusion process. (More precisely, Halfin and Whitt 1981, Theorems 2 and 3 showed the following: If $X^N(0) \Rightarrow \widetilde{X}(0)$, then $X^N(\cdot) \Rightarrow \widetilde{X}(\cdot)$, where $\widetilde{X}(t)$ is a one-dimensional diffusion with infinitesimal drift $m(x)$ given by $m(x) = -\mu\beta$ if $x \geqslant 0$, and $m(x) = -\mu(\beta + x)$ if $x < 0$, and constant infinitesimal variance 2μ.) The steady-state distribution of $\widetilde{X}(\cdot)$ is given by $P(\widetilde{X} > 0) = \alpha$, $P(\widetilde{X} > x \mid \widetilde{X} > 0) = e^{-x\beta}$, $x > 0$, and $P(\widetilde{X} \leqslant x \mid \widetilde{X} \leqslant 0) = \Phi(\beta + x)/\Phi(\beta)$, $x \leqslant 0$. The notation $\Rightarrow$ is used to denote weak convergence in $D[0, \infty)$ (see, e.g., Billingsley 1968, §§14–15), or convergence in distribution. (Hereafter, limiting processes will carry a tilde.)

We conclude with a few informal remarks on the behavior of multiserver systems in the Halfin-Whitt regime. The various assertions that we make here will be rigorously established in §4 where we study the system of original interest.

(1) *Scaling behavior.* The interpretation of the process $X^N(\cdot)$ is as follows: when $X^N(t) > 0$, it is equal to the scaled total number of jobs in both queues, whereas when $X^N(t) < 0$, $-X^N(t)$ is equal to the scaled number of idle servers in the system. This result highlights that for systems with balanced capacity (that is, when the system is neither systematically under-utilized, nor over-utilized) the natural scale that emerges is of order $\sqrt{N}$. Specifically, the total number of customers in the system is approximately $Z^N = N + \sqrt{N}X$, which implies that both queue lengths and the total number of idle servers in the system are of order $\sqrt{N}$.

(2) *Waiting times.* The preceding results imply that the waiting times encountered by customers in both classes will be of order $1/\sqrt{N}$. Intuitively, N busy servers will take $\mathcal{O}(1/\sqrt{N})$ to clear a backlog of $\mathcal{O}(\sqrt{N})$ customers. This observation has an important design implication on the choice of the upper bound for class 2 delay, D_2^N. Specifically, it is natural to assume that D_2^N scales according to

$$D_2^N = \frac{\widetilde{D}_2}{\sqrt{N}} \tag{5}$$

for some appropriate value of $\widetilde{D}_2 > 0$. That is, as the system size grows, the delay guarantee can become tighter and tighter, as prescribed by the natural scaling behavior of the system. Setting the delay guarantee according to (5) is consistent with the order of magnitude of the actual waiting times encountered in the system. The value of $\widetilde{D}_2$ can be chosen to optimize certain system performance measures; this will be done numerically in §5.1.

(3) *Routing policy.* Given (5) and the fact that the arrival rates into each class are of order N, we conclude that for large N, the threshold $\theta^N(t)$ used in our routing policy will be of the form

$$\theta^N = \tilde{\theta}\sqrt{N} \tag{6}$$

for some $\tilde{\theta} > 0$, to be identified later on. That is, the dependence on t is $o(\sqrt{N})$. For any $\theta^N(t)$, this policy is nonidling, and thus satisfies the assumptions and properties mentioned so far.

(4) *Queue-length behavior.* As mentioned above, $\widetilde{X}(t)^+$ represents the total number of customers in queue at time t. (The following conventions are used throughout: $x \wedge y = \min\{x, y\}$, $x \vee y = \max\{x, y\}$, and $x^+ = \max\{x, 0\}$.) Under the threshold routing policy, it is easy to argue that this total queue length will be split into the two classes as follows (see Armony and Maglaras 2004, Proposition 3.1):

$$\widetilde{X}_1(t) = (\widetilde{X}(t) - \tilde{\theta})^+ \quad \text{and} \quad \widetilde{X}_2(t) = \widetilde{X}(t)^+ \wedge \tilde{\theta},$$

where $\widetilde{X}_i$ denotes the limit of the normalized queue-length process $Q_i^N/\sqrt{N}$. To see this note that asymptotically the class 2 queue length cannot exceed $\tilde{\theta}$. Indeed, when $\widetilde{X}_2(t) = \tilde{\theta}$, class 2 gets higher priority. At this time, class 2 jobs are arriving at a rate λ_2^N, and servers are becoming available at a rate $N\mu \gg \lambda_2^N$. In the limit, there is always an available server for each new class 2 arrival, and hence the class 2 queue length will always stay below or at the threshold $\tilde{\theta}$. It follows that if $\widetilde{X}(t)^+ > \tilde{\theta}$, the remaining jobs $(\widetilde{X}(t) - \tilde{\theta})^+$ must be held in queue 1. If $\widetilde{X}(t)^+ < \tilde{\theta}$, then $\widetilde{X}_2(t) < \tilde{\theta}$, and an analogous argument would show that $\widetilde{X}_1(t) = 0$ and $\widetilde{X}_2(t) = \widetilde{X}(t)^+$.

4. Analysis of the System That Announces State-Dependent Information

This section studies the system of original interest that announces state-dependent information to customers and offers them a call-back option. This model introduces two analytical challenges: (i) the routing policy must take into account the state-dependent nature of the arrival streams, and (ii) the system manager needs to be able to compute the expected waiting time for class 1 jobs conditional on the state of the system—this depends on the routing policy employed. This calculation is complex because the current estimate may depend on future arrivals that are themselves state dependent. These issues add considerably to the existing complexity of the static model of Armony and Maglaras (2004). In principle if (i)–(ii) were addressed, one could proceed with a brute force numerical analysis of the associated Markov chain model. While this is theoretically possible, it offers limited insights about the structural properties of the system and is only achievable for relatively small systems (tens of agents). Instead we will build on the results and insights discussed in the previous section and rely on an asymptotic analysis. The numerical solution of

the Markov chain will be used to test the accuracy of our approximate analysis.

Motivated by the discussion of §3, hereafter we make the following assumption:

ASSUMPTION 1. *Balanced capacity*: *We assume that the number of servers is selected in a way that "almost matches" the total potential demand for the system. Specifically,*

$$\Lambda^N_{\text{eff}} = \lambda^N_1(0,0) + \lambda^N_2(0,0) = N\mu - \delta\sqrt{N}\mu \quad \text{for some } \delta \in \mathbf{R}.$$

That is, neglecting the second-order delay in each service class, δ measures the nominal distance from heavy traffic—which may be intentionally negative, making the system slightly under-capacitated.

The routing policy. Section 2 proposed a threshold policy with the time-varying threshold $\theta^N(t) = A^N_2(t) - A^N_2(t - D^N_2)$. Given (5) and assuming optimistically that the class 1 waiting times for a system with state-dependent information are of order $1/\sqrt{N}$, as it was for the system of §3, we get that

$$\lim_{N\to\infty} \frac{\theta^N(t)}{\sqrt{N}} = \lim_{N\to\infty} \frac{A^N_2(t) - A^N_2(t - D^N_2)}{\sqrt{N}} = \tilde{\lambda}_2 \widetilde{D}_2 \quad \forall t,$$

where $\tilde{\lambda}_2 = \lim_{N\to\infty} \lambda^N_2(0, D^N_2)/N$. That is, in large systems the time-varying threshold can be replaced by a static threshold

$$\theta^N = \lambda^N_2(0, D^N_2) D^N_2. \tag{7}$$

For large N, $\theta^N \approx \tilde{\theta}\sqrt{N}$ as in (6) for $\tilde{\theta} = \tilde{\lambda}_2 \widetilde{D}_2$.

Estimating the expected waiting time for class 1 *arrivals.* Given the complexity of computing the actual expected waiting time for class 1 service conditioned on the current state of the system $S(t)$, we will proceed to define a pair of approximate estimators for that quantity. The first approximate estimator follows from the discussion of §2 and is given by

$$\hat{w}^N_1(S(t)) := \frac{Q^N_1(t)}{\lambda^N_1(0, D^N_2)}. \tag{8}$$

Note that we have used $\lambda^N_1(0, D^N_2)$ as a proxy for the associated state-dependent arrival rate. The second estimator is a simplification of (8) that exploits some of the insights given in §3. Specifically, under the threshold routing policy, the class 1 queue length can be approximated by $(Q^N_1(t) + Q^N_2(t) - \theta^N)^+$, which leads to the following estimate for the class 1 waiting time:

$$\check{w}^N_1(S(t)) := \frac{(Q^N_1(t) + Q^N_2(t) - \theta^N)^+}{\lambda^N_1(0, D^N_2)}. \tag{9}$$

Note that this can be rewritten as

$$\check{w}^N_1(S(t)) := \frac{(Z^N_1(t) + Z^N_2(t) - N - \theta^N)^+}{\lambda^N_1(0, D^N_2)},$$

which, in contrast to (8), is only a function of the total number of customers in the system process. This will essentially reduce our study to a one-dimensional analysis.

In the sequel we will analyze the asymptotic behavior of the two-class service system under the threshold routing policy defined through (6) and (7) and the waiting time estimator $\check{w}^N_1(S(t))$. We will show that the two estimates $\hat{w}^N$ and $\check{w}^N$ are asymptotically the same, and that the time-varying threshold $\theta^N(t)/\sqrt{N}$ indeed converges to the time invariant limit $\tilde{\theta}$. Finally, we will justify the choices for the threshold parameter and the waiting time estimator through two appropriate asymptotic criteria, and prove that the routing policy is asymptotically optimal in the sense that it minimizes the waiting time for class 1 customers subject to the constraint that all class 2 customers get served within $\widetilde{D}_2$ time units. Section 4.2 will provide some numerical results illustrating the relative accuracy of the various estimators, as well as the overall accuracy of the asymptotic approximation by comparing it with an exact numerical analysis of the underlying Markov chain.

4.1. Asymptotic Analysis

Recall that the total number of customers in the system process behaves like an $M/M/N$ queue with a state-dependent arrival rate that is equal to the aggregate arrival rate into both classes. Under the estimator $\check{w}^N_1(S)$ that depends on the state of the system only through the total queue-length process, the analysis of the total number of customers in the system process becomes a one-dimensional problem. This can be addressed through a straightforward application of Stone's theorem (Stone 1963). The class-level behavior will be derived by rigorously establishing the state-space collapse foreshadowed in §3. Recall the definition $X^N(t) = (Z^N_1(t) + Z^N_2(t) - N)/\sqrt{N}$.

PROPOSITION 1. *Suppose that Assumption* 1 *holds, the delay bound D^N_2 scales according to* (5)*, and that the system manager announces the waiting time estimate $\check{w}^N$ defined in* (9) *and routes customers according to the threshold policy with θ^N defined in* (7)*. If $X^N(0) \Rightarrow \widetilde{X}(0)$, then $X^N \Rightarrow \widetilde{X}$, where $\widetilde{X}$ is the unique* (*strong*) *solution of the following stochastic differential equation*:

$$d\widetilde{X}(t) = [-\tilde{\delta} + f(\widetilde{X}(t))]\mu\, dt + \sqrt{2\mu}\, dB(t), \tag{10}$$

where B is a standard Brownian motion, $\tilde{\delta} = \delta - \zeta \widetilde{D}_2/\gamma$,

$$f(x) = \begin{cases} \dfrac{\kappa}{\gamma} \dfrac{(x - \tilde{\theta})^+}{\tilde{\lambda}_1(0,0)}, & x \geqslant 0, \\ -x, & x < 0, \end{cases} \tag{11}$$

$$\gamma := \mathbf{P}^\tau(u_1(0,\tau) \cup u_2(0,\tau) \geqslant 0),$$

$$\kappa := \frac{\partial \mathbf{P}^\tau(u_1(w,\tau) \cup u_2(0,\tau) \geqslant 0)}{\partial w}\bigg|_{w=0}, \quad \textit{and}$$

$$\zeta := \frac{\partial \mathbf{P}^\tau(u_1(0,\tau) \cup u_2(d,\tau) \geqslant 0)}{\partial d}\bigg|_{d=0}.$$

The proofs are given in the appendix. This result confirms that, for the system where customers react to state-dependent information, the rationalized or Halfin-Whitt limiting regime emerges as a result of rational customer choice behavior (see Corollary 1), and the natural scale for the number of customers in queue and the number of idle servers is again $\sqrt{N}$.

COROLLARY 1. *Under the assumptions of Proposition* 1, *when the state is* $S^N = (Q_1^N, Q_2^N, Z_1^N + Z_2^N)$,

$$\rho^N(S^N) = 1 - \frac{[\tilde{\delta} - f(X^N)]}{\sqrt{N}} + o\left(\frac{1}{\sqrt{N}}\right).$$

In particular, $\lim_{N\to\infty} \rho^N(S) = 1$, *whenever* X^N *converges to a finite limit.*

The drift of the limiting diffusion, given in (10) and (11), admits a simple interpretation. First, the constant term $\tilde{\delta}$ measures how far is the system from the heavy traffic regime in the absence of congestion for class 1 service, i.e., the aggregate arrival rate into the system when $w_1 = 0$ would be $\lambda_a^N(0, D_2^N) = N\mu - \tilde{\delta}\sqrt{N}\mu + o(\sqrt{N})$. (Note that the term $\zeta\widetilde{D}_2/\gamma$ may be equal to 0 depending on the specifics of the choice model; in fact, this is logical if class 2 corresponds to call-back service. On the other hand, if class 2 represents e-mail service, then it is conceivable that $\zeta \neq 0$ representing the fact that there are some customers that will always choose e-mail over real-time service even if $w_1 = 0$. Our choice model allows for both cases. Also, note that γ may be $\neq 1$ (as in the MNL model), but its actual value just serves as an appropriate normalization constant and is not essential to subsequent analysis.) For the state-dependent drift term $f(\widetilde{X}(t))$ we identify three regions of interest: (a) $\widetilde{X}(t) > \tilde{\theta}$: the waiting time quoted to arriving customers is positive, and this discourages some customers from joining as reflected in the term $(\kappa/\gamma)(x-\tilde{\theta})^+/\tilde{\lambda}_1(0,0)$ (note that $\kappa < 0$). (b) $0 \leqslant \widetilde{X}(t) \leqslant \tilde{\theta}$: the waiting time quoted for class 1 service is 0, all servers are busy, and the resulting drift is simply $\tilde{\delta}$. (c) $\widetilde{X}(t) < 0$: there are $-\widetilde{X}(t)$ idle servers which result in a positive drift contribution $-\mu\widetilde{X}(t)$ to reflect that the departure rate out of the system is less than $N\mu$.

The next result focuses on the behavior of the individual queue-length processes and of the associated waiting times for class 1 and 2 service. In the sequel, $W_i^N(t)$ denotes the virtual waiting time at time t (i.e., this is the time that a virtual class i customer would have to wait if he/she joined class i at time t).

PROPOSITION 2. *Under the assumptions of Proposition* 1 *and that* $(Q_1^N(0)/\sqrt{N}, Q_2^N(0)/\sqrt{N}) \to ((\widetilde{X}(0) - \tilde{\theta})^+, \widetilde{X}(0)^+ \wedge \tilde{\theta})$ *in probability, for every* $t \geqslant 0$, *as* $N \to \infty$,

$$\begin{aligned} \frac{Q_1^N(t)}{\sqrt{N}} &\Rightarrow (\widetilde{X}(t) - \tilde{\theta})^+ =: \widetilde{X}_1(t), \\ \frac{Q_2^N(t)}{\sqrt{N}} &\Rightarrow \widetilde{X}(t)^+ \wedge \tilde{\theta} =: \widetilde{X}_2(t), \end{aligned} \tag{12}$$

and $\sqrt{N}(W_1^N(t), W_2^N(t)) \Rightarrow (\widetilde{W}_1(t), \widetilde{W}_2(t))$, *where*

$$\begin{aligned} \widetilde{W}_1(t) &:= \frac{(\widetilde{X}(t) - \tilde{\theta})^+}{\tilde{\lambda}_1} = \frac{\widetilde{X}_1(t)}{\tilde{\lambda}_1} \quad \textit{and} \\ \widetilde{W}_2(t) &:= \frac{\widetilde{X}(t)^+ \wedge \tilde{\theta}}{\tilde{\lambda}_2} = \frac{\widetilde{X}_2(t)}{\tilde{\lambda}_2}. \end{aligned} \tag{13}$$

Given that $\check{w}_1^N$ and D_2^N are of order $1/\sqrt{N}$, and using the continuity and differentiability assumptions of the choice model, a simple Taylor expansion for $\lambda_2^N(w_1^N, D_2^N)$ gives that $\lambda_2^N(w_1^N, D_2^N) = N\tilde{\lambda}_2 + \sqrt{N}(\cdots) + o(\sqrt{N})$. It now easily follows that $\theta^N(t) = \int_{t-D_2^N}^t dA_2^N(s) = \sqrt{N}\tilde{\theta} + o(\sqrt{N})$. This validates the use of a constant threshold in the routing policy.

Note that (12) establishes that indeed the two waiting time estimators $\hat{w}$ and $\check{w}$ are asymptotically equivalent. Specifically, the appropriately scaled terms for Q_1^N and $(Q_1^N + Q_2^N - \theta^N)^+$ both converge to the common limit $\widetilde{X}_1 = (\widetilde{X} - \tilde{\theta})^+$.

The remainder of this subsection examines the properties of the proposed waiting time estimator and threshold routing policy. First, we define a pair of appropriate asymptotic performance criteria.

DEFINITION 1. Suppose that for every N, $w_1^N(S)$ is the estimated waiting time for a class 1 customer when the state of the system is S. The sequence of estimators $w_1^N(\cdot)$ is called asymptotically consistent for a routing policy π, if for all t,

$$\sqrt{N}[W_1^{N,\pi}(t) - w_1^N(S(t))] \Rightarrow 0 \quad \text{as } N \to \infty, \tag{14}$$

where $W_1^{N,\pi}(t)$ is the actual waiting time experienced by a class 1 customer, who arrives at time t, when the state of the system is $S(t)$.

Roughly speaking an *asymptotically consistent* estimator is one that becomes accurate as the number of servers grows large. The blow-up factor of $\sqrt{N}$ compensates for the fact that the waiting times are decaying to zero like $1/\sqrt{N}$. The second criterion, which was first introduced in Plambeck et al. (2001), addresses the issue of delay constraint qualification for class 2 customers.

DEFINITION 2. A policy π is said to be asymptotically compliant if

$$\left[\sqrt{N}W_2^{N,\pi}(t) - \widetilde{D}_2\right]^+ \Rightarrow 0,$$

where the superscript π denotes the dependence of W_2^N on the policy.

That is, a policy π is asymptotically compliant, if the limit of the appropriately scaled class 2 waiting time always satisfies its delay constraint. The next proposition summarizes the basic properties of the proposed estimator and routing policy.

PROPOSITION 3. *Under the assumptions of Propositions* 1 *and* 2:

(1) *The estimator* $\check{w}_1^N(\cdot)$ *is asymptotically consistent.*

(2) *The threshold routing policy is asymptotically compliant.*

(3) *Fix the waiting time estimator* $\check{w}_1^N$ *and consider any nonidling, nonpreemptive, asymptotically compliant routing policy* π *for which the* (*weak*) *limit queue-length and waiting time processes,* $\widetilde{X}_i^\pi$, $\widetilde{W}_i^\pi$, *exist. Let* π^* *denote the threshold policy with* $\tilde{\theta} = \tilde{\lambda}_2 \widetilde{D}_2$. *Then,* π^* *is asymptotically optimal in the sense that* $\widetilde{W}_1^{\pi^*}(t) \leqslant \widetilde{W}_1^\pi(t)\ \forall t \geqslant 0$ *w.p.*1.

In relation to the last point above, we note that the information signal $\check{w}_1$ will not be asymptotically consistent under arbitrary routing policies π. Thus, the interpretation of point 3 is as follows: Assuming that the system announces $\check{w}$ as its information signal, then π^* is optimal over the set of asymptotically compliant routing policies π in the sense explained above. Ideally, one would like to strengthen the result and allow the system to announce $\hat{w}_1 = Q_1^\pi/\lambda_1$ as its waiting time estimate, which would turn out to be asymptotically consistent for π, and then establish the requisite optimality of π^* in that setting. This result, however, would require a much more involved asymptotic analysis which will not be pursued in this paper. (The complexity is due to the fact that the essentially one-dimensional analysis that suffices under $\check{w}_1$ breaks down under $\hat{w}_1$.) To close the loop, we remind the reader that the (scaled) difference between $\check{w}_1$ and $\hat{w}_1$ is asymptotically negligible under π^*, making this distinction insignificant.

Finally, we derive the steady-state distribution of the limit process $\widetilde{X}$. This is simple because $\widetilde{X}$ is a combination of three well-studied processes: an O-U process for $x \geqslant \tilde{\theta}$, a Brownian motion for $x \in [0, \tilde{\theta}]$, and another O-U process for $x < 0$.

COROLLARY 2. *Let* $b = (-\kappa/\gamma)(1/\tilde{\lambda}_1)$ *and* $\tilde{\delta} = \delta - \zeta \widetilde{D}_2/\gamma$. *If* $\tilde{\delta} \neq 0$, *the p.d.f. of* $\widetilde{X}$ *is*

$$\psi_{\widetilde{X}}(x) = \begin{cases} (1-\alpha_1)\dfrac{\phi(\tilde{\delta}+x)}{\Phi(\tilde{\delta})}, & x \leqslant 0, \\ \alpha_2 \tilde{\delta} e^{-\tilde{\delta}x}, & x \in [0, \tilde{\theta}], \\ \alpha_3 \sqrt{b}\dfrac{\phi(\sqrt{b}(x-\tilde{\theta}) + \tilde{\delta}/\sqrt{b})}{\Phi(-\tilde{\delta}/\sqrt{b})}, & x \geqslant \tilde{\theta}, \end{cases}$$

where the constants $\alpha_1, \alpha_2, \alpha_3 \in [0, 1]$ *and are given by*

$$\alpha_2 = \left[\frac{\tilde{\delta}\Phi(\tilde{\delta})}{\phi(\tilde{\delta})} + 1 - e^{-\tilde{\delta}\tilde{\theta}} + \frac{\tilde{\delta}}{\sqrt{b}} e^{-\tilde{\delta}\tilde{\theta}} \frac{\Phi(-\tilde{\delta}/\sqrt{b})}{\phi(\tilde{\delta}/\sqrt{b})}\right]^{-1},$$

$$\alpha_1 = 1 - \alpha_2 \tilde{\delta}\frac{\Phi(\tilde{\delta})}{\phi(\tilde{\delta})}, \quad \textit{and} \quad \alpha_3 = \frac{\alpha_2 \tilde{\delta}}{\sqrt{b}} e^{-\tilde{\delta}\tilde{\theta}} \frac{\Phi(-\tilde{\delta}/\sqrt{b})}{\phi(\tilde{\delta}/\sqrt{b})}.$$

If $\tilde{\delta} = 0$, $\psi_{\widetilde{X}}(x) = \alpha_2$ *for all* $x \in [0, \tilde{\theta}]$, *and the new* $\alpha_1, \alpha_2, \alpha_3$ *are given in* (27).

4.2. Performance Approximations

This section first demonstrates the accuracy of the various waiting time estimators through a numerical example and then develops a sequence of performance approximations using the steady-state distribution given in Corollary 2. The goal here is to use the limit theory derived above to approximate the performance of an actual system, say with $N = 50$ servers, an MNL choice model with specific parameters, etc.

Quality of the Waiting Time Estimators. Figure 2 compares the "true" expected waiting time for class 1 service as a function of the state (obtained via simulation) with the one predicted using (8) and (9). The parameters of the two systems were selected so that the effective load $\rho_{\text{eff}}(D_2^N) \approx 0.95$. Note that this places the two systems at different operating regimes when measured in the Halfin-Whitt scale expressed in the form $\rho_{\text{eff}}(D_2^N) = 1 - \tilde{\delta}/\sqrt{N}$, but this is not that essential at this stage where the goal was to provide some "numerical verification" of the asymptotic consistency of the two estimators. We make a few additional observations about these plots.

(a) The accuracy of the estimators increases with N and the waiting times decrease as N grows larger, which is consistent with the theory. (Note the difference in the scales of the two plots.) Also, $N = 50$ servers is still very modest in the context of modern-day contact centers.

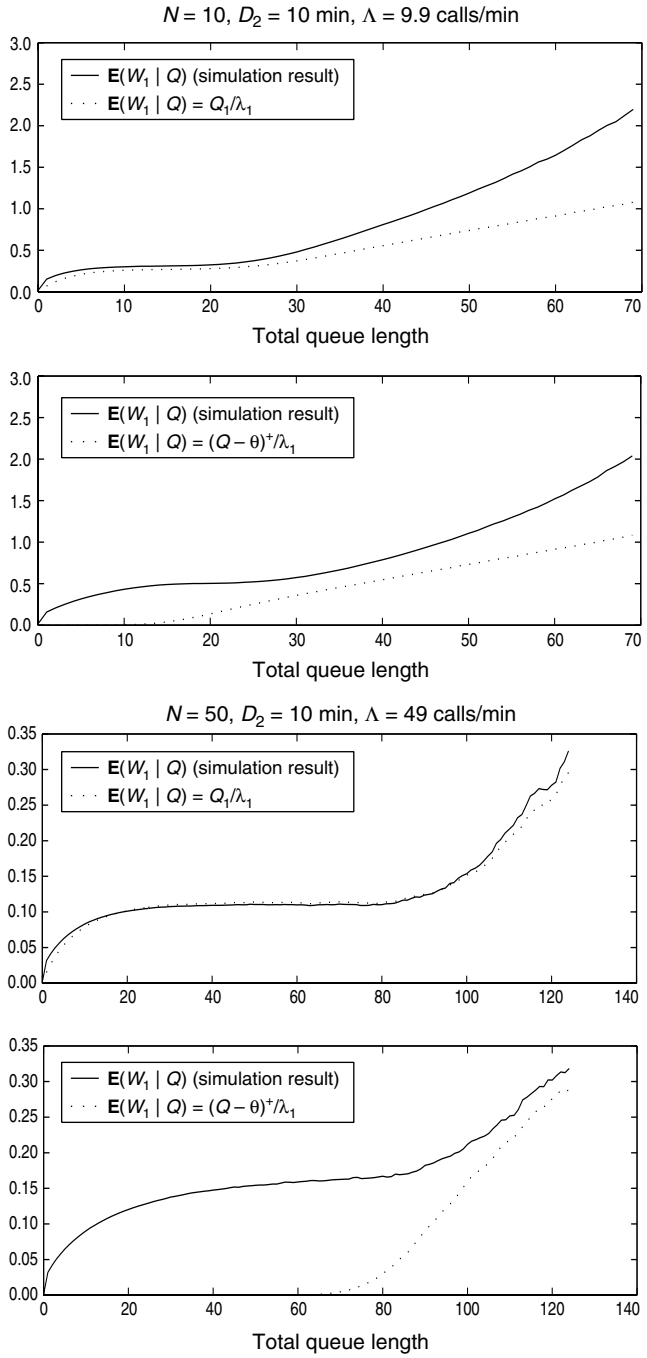
(b) The estimator $\hat{w} = Q_1/\lambda_1$ is more accurate than the one derived using the state-space collapse result $\check{w} = (Q_1 + Q_2 - \theta)^+/\lambda_1$. The difference is in the region where the total queue length $Q = Q_1 + Q_2$ is below the threshold θ and class 1 gets higher priority. For these states, the asymptotic analysis predicts that queue 1 will be empty. In the actual system, however, class 1 jobs will still have to wait for a server to become available, which will translate into a small queue buildup and an associated wait. For example, in the flat part of the curves for $N = 50$ that corresponds to states where $Q \leqslant \theta$, the average class 1 queue length was 2 jobs, which resulted in a small offset from the zero waiting time predicted via the asymptotic analysis. This is not surprising because the appropriate interpretation of the asymptotic result is that the class 1 queue length will be "negligible" in the $\sqrt{N}$ scale of the total queue-length process. This is, indeed, the case and this error disappears as N grows.

(c) The piecewise linear form of the waiting time profile as a function of the total queue length is consistent with the functional form of $\check{w} = (Q - \theta)^+/\lambda_1$.

(d) Actual waiting times experienced by customers will vary around the expected values reported in these plots. A rough calculation shows that when the expected waiting time for large N is close to Q_1/λ_1, the associated standard deviation will be $\sqrt{Q_1}/\lambda_1$ (this is the standard deviation of the sum of Q_1 exponential random variables with rate λ_1), which is of order $N^{-3/4}$.

Accuracy of the Steady-State Distribution Approximation. Figure 3 compares the steady-state distribution given in Corollary 2 with the one computed by exact

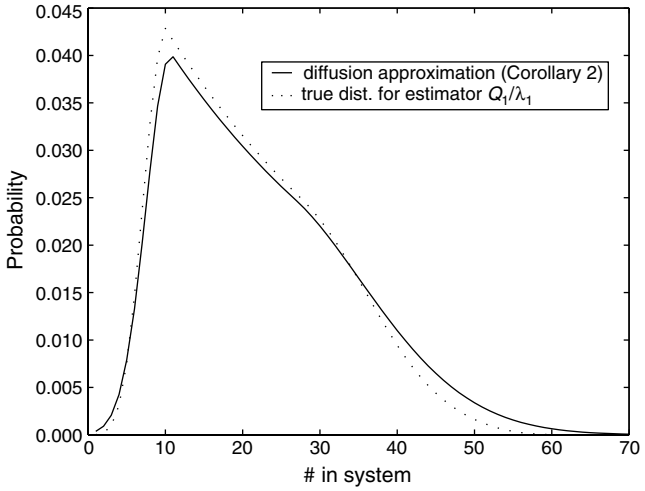
Figure 2. Comparison of W_1 estimators under the MNL choice model.



Note. $r_1 = r_2 = 1$, $c_1 = 0.5$, $c_2 = 0.05$, $\nu = 0.3$, $\mu = 1$, $\theta^{10} = \lambda_2^{10}(0, D_2)D_2 = 15$, $\theta^{50} = \lambda_2^{50}(0, D_2)D_2 = 76$, and λ_1 refers to $\lambda_1^N(0, D_2^N)$ in both plots.

analysis of the associated Markov chain when the system announces $\hat{w}_1(S) = Q_1/\lambda_1(0, D_2^N)$ to arriving customers. The latter was evaluated numerically. To keep the complexity of this calculation low, we studied a system with few ($N = 10$) servers. Even in this small system, the diffusion approximation was quite accurate. Also, it is easy to notice the three different regions identified in the corollary, where

Figure 3. Comparison of steady-state distribution for the total queue-length process for a two-class system with state dependent delay information.



Note. $N = 10$, $\mu = 1$, $D_2 = 10$, $\Lambda = 9.9$, $r_1 = r_2 = 1$, $c_1 = 0.5$, $c_2 = 0.05$, and $\nu = 0.3$.

the total queue-length distribution is normal, exponential, and then normal again. The accuracy of these approximations increases as N grows.

Finally, we use the steady-state distribution of $\widetilde{X}$ to approximate the system performance.

Approximations for the Waiting Time for Real-Time Service in Steady-State.

$$\begin{aligned}\mathbf{E}W_1^N &\approx \frac{\sqrt{N}\mathbf{E}(\widetilde{X}-\tilde{\theta})^+}{\lambda_1^N(0, D_2^N)}\\ &= \frac{\alpha_3\sqrt{N}}{\lambda_1^N(0, D_2^N)}\int_\theta^\infty (x-\tilde{\theta})\sqrt{b}\frac{\phi(\sqrt{b}(x-\tilde{\theta})+\tilde{\delta}/\sqrt{b})}{\Phi(-\tilde{\delta}/\sqrt{b})}\,dx\\ &= \frac{\alpha_3\sqrt{N}}{\lambda_1^N(0, D_2^N)}\left(-\frac{\tilde{\delta}}{b}+\frac{1}{\sqrt{b}}\frac{\phi(\tilde{\delta}/\sqrt{b})}{\Phi(-\tilde{\delta}/\sqrt{b})}\right), \qquad (15)\end{aligned}$$

and for any $x > 0$, $\mathbf{P}(W_1^N > x) \approx \mathbf{P}(\widetilde{W}_1 > \sqrt{N}x)$, which implies that

$$\begin{aligned}&\mathbf{P}(W_1^N > x)\\ &\approx \mathbf{P}(\widetilde{X} > \tilde{\theta} + \lambda_1(0, D_2^N)\sqrt{N}x)\\ &= \frac{\alpha_3}{\Phi(-\tilde{\delta}/\sqrt{b})}\int_{\tilde{\theta}+\lambda_1(0, D_2^N)\sqrt{N}x}^\infty \sqrt{b}\phi\left(\sqrt{b}(x-\tilde{\theta})+\frac{\tilde{\delta}}{\sqrt{b}}\right)dx\\ &= \alpha_3\frac{\Phi(-\tilde{\delta}/\sqrt{b}-\sqrt{b}\lambda_1(0, D_2^N)\sqrt{N}x)}{\Phi(-\tilde{\delta}/\sqrt{b})}. \qquad (16)\end{aligned}$$

Similarly, one can show that

$$\mathrm{Var}(W_1^N) \approx \frac{\alpha_3 N}{b\lambda_1^N(0, D_2^N)^2} - \frac{\tilde{\delta}\sqrt{N}\mathbf{E}W_1^N}{b\lambda_1^N(0, D_2^N)} - (\mathbf{E}W_1^N)^2.$$

The constants b and α_3 are given in Corollary 2.

Total Expected Arrival Rate. This can also be approximated using our asymptotic results. From Proposition 1 we know that the waiting time estimate $\check{w}_1^N(s)$ is of order $\mathcal{O}(1/\sqrt{N})$. Also, from the definition of $\tilde{\delta}$ and $\Lambda_{\text{eff}}^N(D_2^N) = N\mu - \tilde{\delta}\sqrt{N}\mu$, it follows that $\Lambda^N = (N\mu - \tilde{\delta}\sqrt{N}\mu)/\gamma$. Putting these two together, the aggregate arrival rate can be approximated by

$$\begin{aligned}
&\lambda_1^N(\check{w}_1^N(s), D_2^N) + \lambda_2^N(\check{w}_1^N(s), D_2^N)\\
&\quad = \Lambda^N \mathbf{P}^\tau(u_1^\tau(\check{w}_1^N(s)) \cup u_2^\tau(D_2^N) > 0)\\
&\quad \approx \Lambda^N(\gamma + \kappa\check{w}_1^N(s))\\
&\quad \approx N\mu - \sqrt{N}\mu\left(\tilde{\delta} - \frac{\kappa}{\gamma}\sqrt{N}\check{w}_1^N(s)\right).
\end{aligned}$$

Taking expectations with respect to the steady-state distribution given in Corollary 2 we may approximate the total expected arrival rate into the system by

$$\begin{aligned}
&\mathbf{E}[\lambda_1^N(\check{w}_1^N(s), D_2^N) + \lambda_2^N(\check{w}_1^N(s), D_2^N)]\\
&\quad \approx N\mu - \sqrt{N}\mu\left(\tilde{\delta} - \frac{\kappa}{\gamma}\sqrt{N}\mathbf{E}\check{w}_1^N\right)\\
&\quad \approx N\mu - \sqrt{N}\mu\left(\tilde{\delta} - \frac{\kappa}{\gamma}\sqrt{N}\mathbf{E}W_1^N\right). \qquad (17)
\end{aligned}$$

Also, the asymptotic loss in throughput due to congestion is

$$\Lambda_{\text{eff}}^N(D_2^N) - \mathbf{E}\lambda_a^N(D_2^N) = \sqrt{N}\mu\frac{\kappa}{\gamma}\mathbf{E}\widetilde{W}_1. \qquad (18)$$

Similarly, $\mathbf{E}W_2^N \approx (\sqrt{N}\mathbf{E}\widetilde{X}^+ \wedge \tilde{\theta})/(\lambda_2^N(0, D_2^N))$, which, after some simple manipulations using the distribution in Corollary 2, gives

$$\mathbf{E}W_2^N \approx \frac{\alpha_2\sqrt{N}}{\lambda_2^N(0, D_2^N)}\left[\frac{1}{\tilde{\delta}} - e^{-\tilde{\delta}\tilde{\theta}}\left(\tilde{\theta} + \frac{1}{\tilde{\delta}}\right)\right]. \qquad (19)$$

Related expressions can be obtained for the corresponding variance and other quantities of interest. Using these closed-form characterizations, §5.1 will study the dependence of the system behavior on important model parameters.

5. Performance Improvements and Staffing Rules

This section uses the closed form performance characterizations derived above to study three practical questions: (a) Does announcing state-dependent information help? (b) What is the value of the call-back option, and how should the system manager select the delay guarantee to optimize performance? (c) How many servers should the system have to satisfy the typical operational specifications imposed in call centers (e.g., average waiting time $\leqslant 10$ seconds, 80% of calls answered within 20 seconds, etc.)?

5.1. The Value of Announcing State-Dependent Information

First, we compare the performance of the system analyzed so far to the one that announces the steady-state expected waiting time for class 1 service instead, and illustrate how providing the state-dependent information improves overall system performance. This "static" case was studied in Armony and Maglaras (2004), the main results of which are summarized below.

5.1.1. Background: Analysis of the System That Announces Steady-State Delay Information. In this case, customers are assumed to make decisions based on the steady-state expected waiting time for class 1 service and the call-back deadline. This information is either announced or is assumed to be known by the customers after repeated visits to the system. This model requires a statistical equilibrium analysis: one needs to find the waiting time estimate whose announcement results in arrival rates that are consistent with the steady-state expected waiting time quoted in the first place.

In the sequel a superscript "s" will mark all processes associated with this "static" system. To start with, the structural insights of §3 are still in force, the system still operates under a threshold routing policy with $\theta^N = \lambda_2^N(0, D_2^N)D_2^N$ (same as before), but announces $\mathbf{E}\check{w}_1^{s,N}$ in place of $\check{w}_1^N(S(t))$. This results in a change in the infinitesimal drift of the limit of the scaled total number of customers in the system process, which is now governed by the following stochastic differential equation:

$$\begin{aligned}
d\widetilde{X}^s(t) = &\left[\tilde{\delta} - \frac{\kappa}{\gamma}\frac{\mathbf{E}(\widetilde{X}^s - \theta)^+}{\tilde{\lambda}_1} + \min(\widetilde{X}^s(t), 0)\right]\mu\, dt\\
&+ \sqrt{2\mu}\, dB(t),
\end{aligned}$$

where $\mathbf{E}\widetilde{W}_1^s = (\mathbf{E}(\widetilde{X}^s - \theta)^+)/\tilde{\lambda}_1$. This equation makes the change from (10) transparent. This is the same as the Halfin-Whitt diffusion (described in §3), where the constant part of the drift, denoted here by $\beta(\mathbf{E}\widetilde{W}_1^s, \widetilde{D}_2)$, is given by

$$\beta(\mathbf{E}\widetilde{W}_1^s, \widetilde{D}_2) = \tilde{\delta} - \frac{\kappa}{\gamma}\mathbf{E}\widetilde{W}_1^s, \qquad (20)$$

where γ, κ, and $\tilde{\delta}$ are the same constants defined in §4. Using the steady-state distribution associated with $\widetilde{X}^s$ that was given in §3, we conclude that the system will be in statistical equilibrium if and only if

$$\mathbf{E}\widetilde{W}_1^s = \frac{1}{\tilde{\lambda}_1}\frac{\alpha(\beta(\mathbf{E}\widetilde{W}_1^s, \widetilde{D}_2))}{\beta(\mathbf{E}\widetilde{W}_1^s, \widetilde{D}_2)}e^{-\beta(\mathbf{E}\widetilde{W}_1^s, \widetilde{D}_2)\tilde{\theta}}, \qquad (21)$$

where $\alpha(\beta) = \mathbf{P}(\widetilde{X} \geqslant 0) = [1 + \sqrt{2\pi}\beta\Phi(\beta)e^{\beta^2/2}]^{-1}$ (defined in §3). Expressions (20) and (21) define a fixed point equation for $\mathbf{E}\widetilde{W}_1^s$. Proposition 4.1 in Armony and Maglaras (2004) established that this equation admits a unique solution $\mathbf{E}\widetilde{W}_1^s$ that characterizes the unique and stable equilibrium point of the limiting system.

Performance approximations. Let β^* denote this equilibrium parameter, and set $\alpha^* = \alpha(\beta^*)$. Then, using the steady-state distribution of $\widetilde{X}^s$ and an appropriate variant of Proposition 2 (Armony and Maglaras 2004, Propositions 3.2–3.3), we can derive closed form expressions for several quantities of interest. First, note that $\lambda_a^{s,N} = N\mu - \beta^*\sqrt{N}\mu + o(\sqrt{N})$. As further examples,

$$\mathbf{E}W_1^{s,N} \approx \frac{1}{\lambda_1^{s,N}} \frac{\alpha^*\sqrt{N}}{\beta^*} e^{-\beta^*\tilde{\theta}} \tag{22}$$

and

$$\mathbf{P}(W_1^{s,N} > y) \approx \mathbf{P}(\widetilde{W}_1^s > y\sqrt{N}) = \alpha^* e^{-\beta^*(\tilde{\theta}+\tilde{\lambda}_1\sqrt{N}y)}.$$

5.1.2. Numerical Comparisons. In this section, the dynamic and the static systems are compared by analyzing their asymptotic performance. The next proposition shows that, for the same delay bound D_2^N, the asymptotic performance of the dynamic system dominates that of the static one.

Let $\widetilde{X}^d$, $\widetilde{W}_1^d$, $\widetilde{X}^s$, and $\widetilde{W}_1^s$ be the limits for the total queue-length and class 1 waiting time processes in the dynamic and static systems defined in Proposition 1 and §5.1.1, respectively. To simplify calculations, the next result focuses on the case $\tilde{\delta} = \delta - \zeta\widetilde{D}_2/\gamma = 0$.

PROPOSITION 4. *If* $\tilde{\delta} = 0$, *then*

$$\mathbf{E}\widetilde{W}_1^d \leqslant \mathbf{E}\widetilde{W}_1^s.$$

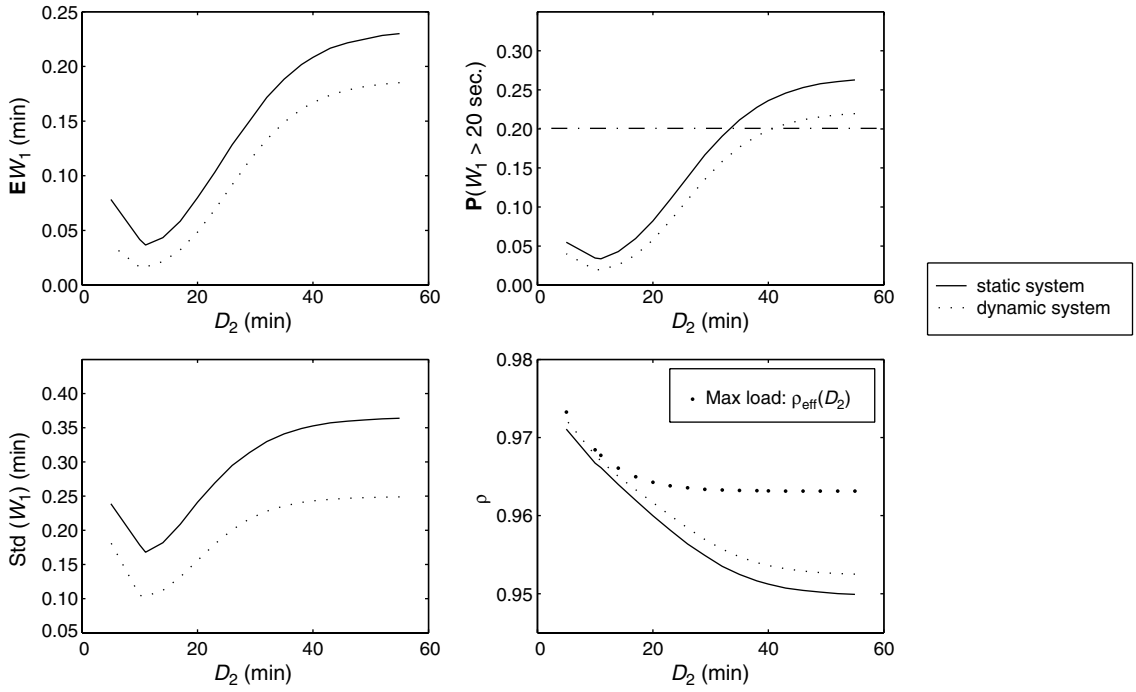
Assuming that for each N the static system has a unique equilibrium point characterized by its steady-state expected waiting time $\mathbf{E}W_1^{s,N}$, *then*

$$\lim_{N\to\infty} \frac{1}{\sqrt{N}}[\mathbf{E}\lambda_a^{N,d}(W_1^{d,N}, D_2^N) - \lambda_a^{N,s}(\mathbf{E}W_1^{s,N}, D_2^N)] \geqslant 0.$$

That is, dynamic information reduces the expected waiting time for class 1 service, while increasing the aggregate traffic that gets served by the system! Whitt (1999a) derived a similar result for a single class $M/M/N$ system with reneging. (We suspect the same line of proof will work for the case $\tilde{\delta} \neq 0$. However, due to the more complex nature of the steady-state distribution of $\widetilde{X}^d$ in Corollary 2 the algebraic manipulations become messier and we have not been able to resolve this issue. The key difference is in the constant α_3 defined in the corollary if $\tilde{\delta} \neq 0$ or in (27) if $\tilde{\delta} = 0$. The results presented in this section will provide numerical evidence in support of the claim for the general case when $\tilde{\delta} \neq 0$.)

Figure 4 compares the performance of the static and dynamic systems for a scenario where the total effective traffic into the system is slightly below capacity ($\rho_{\text{eff}}(0) \triangleq \Lambda_{\text{eff}}(0)/(N\mu) = 0.98$). The figures are computed using our asymptotic approximations. In all four comparisons the dynamic system outperforms the static one. For any fixed value of D_2, the expected waiting time for class 1 service was reduced by 20%–50%. The probability that the waiting time in class 1 exceeds 20 seconds (a typical specification for call centers) is decreased by 0.015 to 0.04.

Figure 4. Comparison of $\mathbf{E}W_1^N$, $\mathbf{P}(W_1^N > 20$ sec.), Std(W_1^N), and ρ under steady-state (static) and state-dependent (dynamic) information.



Note. System: $N = 50$, $\mu = 1$, MNL choice model with $r_1 = r_2 = 1$, $c_1 = 0.5$, $c_2 = 0.05$, $\nu = 0.3$, and $\rho_{\text{eff}}(0) = 0.98$.

Table 1. Sensitivity w.r.t. the maximum offered load $\rho_{\text{eff}}(0)$.

$\rho_{\text{eff}}(0)$	$(\mathbf{E}W_1^s, \mathbf{P}(W_1^s > 20s.))$	$(\mathbf{E}W_1^d, \mathbf{P}(W_1^d > 20s.))$	$(D_2^s, \mathbf{E}W_2^s)$	$(D_2^d, \mathbf{E}W_2^d)$	ρ^s	ρ^d
0.95	(0.002, 0.002)	(0.001, 0.001)	(8, 0.87)	(9, 0.98)	0.940	0.940
0.975	(0.024, 0.022)	(0.011, 0.013)	(10, 1.89)	(10, 1.95)	0.963	0.963
1.00	(0.185, 0.132)	(0.080, 0.094)	(13, 3.13)	(12, 3.50)	0.978	0.983
1.025	(0.502, 0.313)	(0.307, 0.349)	(18, 3.52)	(17, 3.79)	0.983	0.994

Note. System: $N = 50$, $\mu = 1$, MNL choice model with $r_1 = r_2 = 1$, $c_1 = 0.5$, $c_2 = 0.05$, and $\nu = 0.3$. Waiting times reported in minutes.

The performance improvement with respect to the aggregate traffic into the system is very small, but this is expected because both systems are operating very efficiently (i.e., with small expected waiting times in class 1), while being close to the heavy traffic regime.

In the sequel, we perform a sequence of numerical experiments that investigate the relative performance of the two systems as we vary some of the model parameters. The base case is a system with $N = 50$ servers and $\mu = 1$, and the MNL choice model with $r_1 = r_2 = 1$, $c_1 = 0.5$, $c_2 = 0.05$, and $\nu = 0.3$. For each set of model parameters we will only report results that correspond to the (optimal) delay bound for class 2 service that minimizes the expected waiting time for class 1. (This is selected numerically after first evaluating the system performance at different values of D_2 using the asymptotic approximations.) While the numerical values reported below depend strongly on the nature of the choice model, the sensitivity results are more robust and illustrate the effect of state-dependent information as well as the value of the call-back channel.

(i) *Dependence on the extent to which capacity is balanced* (δ). Our first test examines the comparative advantage of state-dependent over steady-state information as we vary the effective load into the system as measured by $\rho_{\text{eff}}(0)$. Recall that $\rho_{\text{eff}}(0)$ is equal to the maximum possible load for the system that corresponds to the case where $W_1 = D_2 = 0$. So, the quantity $1 - \rho_{\text{eff}}$ measures the nominal excess capacity of the system.

The results of Table 1 illustrate that the performance gains due to state-dependent information are significant for a wide range of load factors. As expected, when the offered load increases, the waiting times also increase. While the difference in throughput rate is small, as both systems are very efficient, the dynamic one maintains a small advantage. These results lend credibility to the result of Proposition 4 in the case where $\tilde{\delta} \neq 0$. (The value of $\tilde{\delta}$ in these experiments was ranging from 0.43 to -0.05 as the load was increasing.) The performance under state-dependent information seems a lot more robust when the system is more heavily loaded ($\rho_{\text{eff}}(0) \geqslant 1$), because the "active" load balancing when queue 1 grows large is much more efficient. Also, note that while dynamic information leads to significant improvements of class 1's performance, it may degrade class 2's service. However, in the dynamic system class 2's expected delay is still significantly lower than its performance guarantee, and the impact of the class 2 service degradation on overall utility is small because customers tend to be less sensitive to class 2 delay. Finally, we note that the static system dominates the dynamic one in terms of $\mathbf{P}(W_1 > 20s.)$ when $\rho_{\text{eff}}(0) = 1.025$, but this is due to the fact that in this highly congested case the target value of 20 seconds was comparable to the actual expected waiting times in class 1. In practice, this probabilistic constraint is meant to bound the tail behavior of the system (i.e., the target value is well above the expected waiting time), and in this parameter regime the dynamic system was observed to be better.

(ii) *Dependence on the delay sensitivity parameters.* In the experiments reported in Table 2 we studied the behavior of the two systems for different values of the delay sensitivity parameters for the two service options c_1, c_2 that appear in the MNL model. The dynamic system consistently outperforms the static one. As expected, as the delay sensitivity to the class 2 deadline increases (c_2 grows from 0.025 to 0.1), the optimal delay bound D_2 and the expected waiting time for class 2 customers decrease. Similar observations hold true when the delay sensitivity to class 1 service (c_1) is allowed to change.

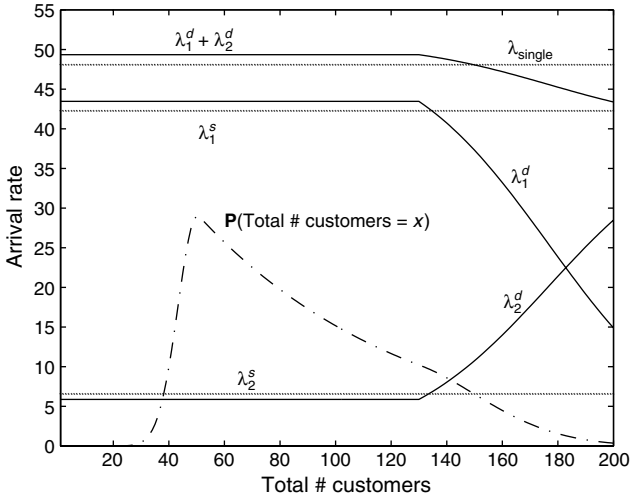
(iii) *Load balancing and the value of the call-back option.* It is evident from Figure 4 that the addition of the call-back (or e-mail) option, together with careful selection of the promised delay bound D_2, can lead to significant

Table 2. Sensitivity w.r.t. delay sensitivity parameters c_1, c_2.

(c_1, c_2)	$\mathbf{E}W_1^s$	$\mathbf{E}W_1^d$	$(D_2^s, \mathbf{E}W_2^s)$	$(D_2^d, \mathbf{E}W_2^d)$	$(\lambda_1^s, \lambda_2^s)$	$(\lambda_1^d, \lambda_2^d)$
(0.5, 0.1)	0.271	0.153	(7, 1.34)	(7, 1.54)	(42.1, 6.4)	(43.5, 5.4)
(0.5, 0.025)	0.093	0.027	(24, 5.86)	(23, 6.66)	(42.4, 6.7)	(42.7, 6.6)
(0.5, 0.05)	0.185	0.080	(13, 3.13)	(12, 3.50)	(42.3, 6.6)	(42.6, 6.6)
(0.25, 0.05)	0.255	0.128	(13, 3.13)	(12, 3.34)	(42.9, 6.1)	(42.8, 6.4)
(1, 0.05)	0.127	0.048	(13, 3.07)	(12, 3.64)	(41.4, 7.2)	(42.4, 6.7)

Note. System: $N = 50$, $\mu = 1$, $\rho_{\text{eff}}(0) = 1$, MNL choice model with $r_1 = r_2 = 1$, and $\nu = 0.3$. Waiting times reported in minutes.

Figure 5. Load balancing via the call-back option and state-dependent waiting-time information.



Note. System: $N = 50$, $\mu = 1$, MNL choice model with $r_1 = r_2 = 1$, $c_1 = 0.5$, $c_2 = 0.05$, $\nu = 0.3$, and $\rho_{\text{eff}}(0) = 1.00$. From Table 1, $D_2 = 12$ and $\theta^N = 80$.

performance improvements. In the sequel, we provide some additional results comparing the performance of the optimized two-class system with a single-class system, which in our setting corresponds to the case $D_2 \to \infty$.

Overall, the introduction of the call-back option with an optimized deadline D_2 resulted in significant performance improvements. The expected waiting time in class 1 gets reduced by a factor of 4 to 10, the probability that the waiting time in class 1 will exceed the target of 20 seconds is reduced by 0.1 to 0.25, while the total throughput rate increased by 1%–2.5%! The margin of improvement reduces as the nominal load increases. The performance improvements, especially in the dynamic system, are due to the efficient load balancing that is achieved through the customers' reaction to the state-dependent information when the system is congested. Figure 5 illustrates this point by contrasting the state-dependent arrival rates for the two-class system with the ones that correspond to the system with static information with or without the call-back option. First, we note that the addition of the call-back option has the effect of shifting some demand from the real-time to the call-back option, in a way that increases the aggregate throughput rate. Second, the system with state-dependent information has a higher aggregate throughput rate in most possible states (i.e., provided that the total number of customers is $\leqslant N + \theta^N = 130$), and actively controls congestion by shifting demand to the call-back channel as the waiting time in class 1 increases. Note that using the steady-state distribution of Corollary 2 we conclude that $\mathbf{P}(\widehat{W}_1^d = 0) = \mathbf{P}(\text{total \# customers} \leqslant N + \theta^N) = 0.83$, and that $\mathbf{P}(\lambda_1^d(S(t)) + \lambda_2^d(S(t)) \geqslant \lambda_1^s + \lambda_2^s) = 0.88$.

5.2. Staffing Rules

A dual point of view on the results of Table 3 is in the context of server staffing. Specifically, consider the problem (see Armony and Maglaras 2004, §5.2) that involves choosing the minimum number of servers to satisfy a set of performance specifications that are typically encountered in contact centers such as:

- the expected waiting time for real-time calls $\leqslant$ 10 seconds,
- 80% of all calls are answered within 20 seconds,
- $\leqslant$1% balking probability.

For example, the results of Table 3 illustrate that the two-class service system with state-dependent information and $N = 50$ servers provides a similar quality of service as the single-class system that announces steady-state waiting-time information, while processing 6.5% higher demand. Conversely, numerical experiments show that the two-class system with dynamic information requires about 5% less capacity to provide similar quality of service to that of a single-class static system, and about 1%–2% less than the two-class static system. Because staffing costs are a dominant consideration in call centers, these performance improvements can be rather significant.

In the sequel we provide a detailed mathematical investigation of the staffing problem. Consider the following problem formulation:

$$\min\{N : \mathbf{E}W_1^N \leqslant w_e,\ \mathbf{P}(W_1^N \geqslant y) \leqslant \epsilon_1,\ \mathbf{P}(\text{balking}) \leqslant \epsilon_b\}, \tag{23}$$

with typical values for these specifications given above. For a fixed D_2, let $R = \Lambda_{\text{eff}}(D_2)/\mu$. The remainder of this section outlines a procedure that culminates in a simple numerical recipe that provides a solution to the staffing problem of the form $N = R + x\sqrt{R}$. This reenforces the well-known square-root staffing rule (e.g., Borst et al. 2004), that, to the best of our knowledge, is validated here for the first time (together with Armony and Maglaras 2004) in a multiclass setting.

Table 3. The value of the call-back option.

$\rho_{\text{eff}}(0)$	Single class ($D_2 = \infty$) ($\mathbf{E}W_1^s$, $\mathbf{P}(W_1^s > 20)$, λ_1^s)	Two classes (opt. D_2^s) ($\mathbf{E}W_1^s$, $\mathbf{P}(W_1^s > 20)$, λ_1^s, λ_2^s)	Single class ($D_2 = \infty$) ($\mathbf{E}W_1^d$, $\mathbf{P}(W_1^d > 20)$, λ_1^d)	Two classes (opt. D_2^d) ($\mathbf{E}W_1^d$, $\mathbf{P}(W_1^d > 20)$, λ_1^d, λ_2^d)
0.95	(0.13, 0.14, 46.3)	(0.002, 0.002, 37.2, 9.8)	(0.11, 0.11, 46.4)	(0.001, 0.001, 38.4, 8.6)
0.975	(0.22, 0.24, 47.3)	(0.02, 0.02, 40.2, 7.9)	(0.17, 0.20, 47.4)	(0.01, 0.01, 40.4, 7.8)
1.00	(0.37, 0.37, 48.1)	(0.19, 0.13, 42.3, 6.6)	(0.27, 0.34, 48.4)	(0.08, 0.09, 42.6, 6.6)

Note. This table contrasts the results reported in Table 1 with the behavior of the associated single-class systems ($D_2 = \infty$). System: $N = 50$, $\mu = 1$, MNL choice model with $r_1 = r_2 = 1$, $c_1 = 0.5$, $c_2 = 0.05$, $\nu = 0.3$, and $\rho_{\text{eff}}(0) \in [0.95, 1.0]$.

Bounds on expected waiting time of the form $\mathbf{E}W_1^N \leqslant w_e$ can be approximated by the specification $\mathbf{E}\widetilde{W}_1 \leqslant \sqrt{R}w_e$. Note that, given the parameters of the choice model and D_2, the value of $\mathbf{E}\widetilde{W}_1$ is only a function of $\tilde{\delta}$ (see (15)). Hence,

$$\mathbf{E}\widetilde{W}_1 \leqslant \sqrt{R}w_e \quad \Rightarrow \quad \tilde{\delta} \geqslant \tilde{\delta}_e(w_e) \triangleq \min\{\tilde{\delta}\colon \mathbf{E}\widetilde{W}_1 \leqslant \sqrt{R}w_e\}.$$

Probabilistic constraints on waiting time of the form $\mathbf{P}(W_1^N \geqslant y) \leqslant \epsilon_1$ (typical parameters are $y = 20$ sec. and $\epsilon_1 = 0.2$) are approximated as follows:

$$\mathbf{P}(W_1^N \geqslant y) \approx \mathbf{P}(\widetilde{W}_1 \geqslant y\sqrt{R}) \leqslant \epsilon_1$$
$$\Rightarrow \quad \tilde{\delta} \geqslant \tilde{\delta}_p(y, \epsilon_1) \triangleq \min\{\tilde{\delta}\colon \mathbf{P}(\widetilde{W}_1 \geqslant y\sqrt{R}) \leqslant \epsilon_1\},$$

where the last step follows again from the fact that $\mathbf{P}(\widetilde{W}_1 \geqslant y\sqrt{R})$ is a function of $\tilde{\delta}$ alone (see (16)).

Bounds on balking of the form $\mathbf{P}(\text{balking}) \leqslant \epsilon_b$ are incorporated using (17) and (18) by noting that

$$\mathbf{P}(\text{balking}) = 1 - \frac{\mathbf{E}\lambda_a^N}{\Lambda_{\text{eff}}(D_2)} \approx -\frac{\kappa}{\gamma}\mathbf{E}W_1^N \leqslant \epsilon_b \quad \Rightarrow \quad \mathbf{E}W_1^N \leqslant -\frac{\gamma}{\kappa}\epsilon_b,$$

which implies that the constraint $\mathbf{P}(\text{balking}) \leqslant \epsilon_b$ can be replaced by

$$\tilde{\delta} \geqslant \tilde{\delta}_e\left(-\frac{\gamma}{\kappa}\epsilon_b\right).$$

Putting it all together, to satisfy the specifications $\mathbf{E}W_1^N \leqslant w_e$, $\mathbf{P}(W_1^N \geqslant y) \leqslant \epsilon_1$, and $\mathbf{P}(\text{balking}) \leqslant \epsilon_b$, one must set $N = R + x^*\sqrt{R}$, where

$$x^* = \max\left(\tilde{\delta}_e(w_e), \tilde{\delta}_p(y, \epsilon_1), \tilde{\delta}_e\left(-\frac{\gamma}{\kappa}\epsilon_b\right)\right),$$

and its value is computed numerically using the steady-state distribution given in Corollary 2 and the expressions given in §4.2.

6. Concluding Remarks

This paper analyzed a two-channel service system, where one channel offers real-time service, and the other offers service within a guaranteed upper bound on delay—such as a call-back or e-mail option. Customer behavior was captured through a probabilistic choice model. An asymptotic analysis was used to develop a near-optimal routing rule, a consistent waiting time estimator, and to obtain analytic approximations for the system's steady-state behavior. This mode of analysis is accurate for systems with many servers that are not significantly over- or under-capacitated, which is arguably the canonical operating regime for such systems.

The key findings are that service systems can improve their performance substantially by (a) offering such call-back or e-mail options with performance guarantees, and (b) by informing their customers of state information upon their arrival. These two initiatives encourage active congestion control and load balancing between the two service options.

Several interesting areas of future research arise. One important extension from a practical viewpoint is to allow for nonstationary arrivals. Other interesting directions lie in the management of the delay bounds for the call-back or e-mail options. For example, the system manager could dynamically vary the promised guarantee depending on the current level of congestion, or allow the customers to make "appointments" in future times where they wish to receive their call-back. In addition, one would want to extend the results of this paper to the case where the service rates for the two classes are not equal (see Maglaras and Zeevi 2004 for the analytical tools required for this extension). Also, from a practical viewpoint, it is often the case that separate pools of servers handle inbound calls or initiate outbound (call-back) calls or reply to e-mails. This paper provides a deeper understanding into the benefits of cross-training agents to perform these tasks.

On the methodological level, two concepts were introduced here that we believe to be general enough to apply to other situations. The first is the notion of *asymptotic consistency*. This may prove useful in formulation and analysis of similar dynamic estimation problems where the random quantity of interest depends on future events, which themselves depend on the value of this quantity in the future. The second useful result is that the state-space of our limiting system collapses to one dimension. We use this result to find estimates of the waiting times in a system with state-dependent arrival rates, for a direct derivation of the waiting times may be impossible. This may be used in other problems in operations management, such as the issue of dynamic pricing of congestion sensitive services.

Appendix. Proofs

Proof of Proposition 1. The proof is based on Stone's theorem (Stone 1963). To see that we first set up our system in the terminology of Iglehart (1965). Specifically, let $Z_a^N(t) = Z_1^N(t) + Z_2^N(t)$ be the total number of jobs in the Nth system at time t, and let $X^N(t) = (Z_a^N(t) - N)/\sqrt{N}$ be the scaled state. $X^N(t)$ is a birth-death process with state-space

$$E^N = \left\{\frac{0-N}{\sqrt{N}}, \frac{1-N}{\sqrt{N}}, \ldots, \frac{N-N}{\sqrt{N}}, \frac{N+1-N}{\sqrt{N}}, \ldots\right\},$$

birth rates

$$\lambda^N\left(\frac{z-N}{\sqrt{N}}\right) = \Lambda^N\mathbf{P}^\tau\left(u_1(\check{w}_1^N(z), \tau) \vee u_2(D_2^N, \tau) \geqslant 0\right) = \Lambda^N\mathbf{P}^\tau\left(u_1\left(\frac{[z-N-\theta^N]^+}{N\tilde{\lambda}_1}, \tau\right) \vee u_2(D_2^N, \tau) \geqslant 0\right),$$

and death rates

$$\mu^N\left(\frac{z-N}{\sqrt{N}}\right)=(z\wedge N)\mu.$$

For any number y, $-\sqrt{N}<y<\infty$, define $z^N(y)$ to be the largest integer such that $(z^N(y)-N)/\sqrt{N}$ is less than or equal to y. Namely,

$$z^N(y)=\max\left\{z\,\middle|\,\frac{z-N}{\sqrt{N}}\leqslant y \text{ and } \frac{z-N}{\sqrt{N}}\in E^N\right\}=\lfloor y\sqrt{N}\rfloor+N,$$

where $\lfloor x\rfloor$ is the largest integer that is less than or equal to x. Given this setup, we can now define the infinitesimal mean of the process $X^N(t)$ to be

$$\begin{aligned}m^N(y)&=\frac{1}{\sqrt{N}}\left[\lambda^N\left(\frac{z^N(y)-N}{\sqrt{N}}\right)-\mu^N\left(\frac{z^N(y)-N}{\sqrt{N}}\right)\right]\\&=\frac{1}{\sqrt{N}}\left[\lambda^N\left(\frac{\lfloor y\sqrt{N}\rfloor}{\sqrt{N}}\right)-\mu^N\left(\frac{\lfloor y\sqrt{N}\rfloor}{\sqrt{N}}\right)\right].\end{aligned}\tag{24}$$

Similarly, the infinitesimal variance is

$$\begin{aligned}(\sigma^2)^N(y)&=\frac{1}{N}\left[\lambda^N\left(\frac{z^N(y)-N}{\sqrt{N}}\right)+\mu^N\left(\frac{z^N(y)-N}{\sqrt{N}}\right)\right]\\&=\frac{1}{N}\left[\lambda^N\left(\frac{\lfloor y\sqrt{N}\rfloor}{\sqrt{N}}\right)+\mu^N\left(\frac{\lfloor y\sqrt{N}\rfloor}{\sqrt{N}}\right)\right].\end{aligned}\tag{25}$$

Define the function $g(x_1,x_2)\triangleq\mathbf{P}^\tau(u_1(x_1,\tau)\vee u_2(x_2,\tau)\geqslant 0)$, and recall that the effective total arrival rate into the system satisfies $\Lambda^N_{\text{eff}}=\Lambda^N\mathbf{P}^\tau(u_1(0,\tau)\vee u_2(0,\tau)\geqslant 0)=\mu N-\mu\sqrt{N}\delta$. Hence, the total potential arrival rate Λ^N may be written as $\Lambda^N=(\mu N-\mu\sqrt{N}\delta)/g(0,0)$. For $i=1,2$, denote by $g'_{x_i}(\cdot,\cdot)=\partial g(\cdot,\cdot)/\partial x_i$.

Now let

$$\begin{aligned}m^0(y)&=-\delta\mu+\frac{g'_{x_2}(0,0)}{g(0,0)}\widetilde{D}_2\mu\\&\quad+\begin{cases}\mu\dfrac{g'_{x_1}(0,0)}{g(0,0)}\dfrac{(y-\tilde\theta)^+}{\tilde\lambda_1}, & y\geqslant 0,\\ -\mu y, & y<0,\end{cases}\end{aligned}$$

and

$$(\sigma^2)^0(y)=2\mu.$$

To prove the proposition, we need to show that Stone's criteria hold. Specifically, we need to verify that

(1) E^N becomes dense in $(-\infty,\infty)$ as $N\to\infty$.

(2) $m^0(\cdot)$ and $(\sigma^2)^0(\cdot)$ are continuous and $(\sigma^2)^0(\cdot)>0$.

(3) $m^N(y)\to m^0(y)$ and $(\sigma^2)^N(y)\to(\sigma^2)^0$ uniformly on compact intervals (u.o.c), as $N\to\infty$.

Showing (1) and (2) is straightforward. We show that (3) holds for all $y>\tilde\theta$. The proof for $y\leqslant\tilde\theta$ is similar (and simpler) and is essentially identical to the proof of the result of Halfin and Whitt, which was presented in §3 (and was proved in Halfin and Whitt 1981).

Let $y>\tilde\theta$, then for N large enough $\lfloor y\sqrt{N}\rfloor>\theta^N=\tilde\theta\sqrt{N}$ and $\lfloor y\sqrt{N}\rfloor+N>N$. Hence,

$$\begin{aligned}&\lambda^N\left(\frac{\lfloor y\sqrt{N}\rfloor}{\sqrt{N}}\right)\\&\quad=\Lambda^N\mathbf{P}^\tau\left(u_1\left(\frac{\lfloor y\sqrt{N}\rfloor-\theta^N}{N\tilde\lambda_1},\tau\right)\vee u_2(D_2^N,\tau)\geqslant 0\right)\\&\quad=\frac{\mu N-\mu\sqrt{N}\delta}{g(0,0)}g\left(\frac{\lfloor y\sqrt{N}\rfloor-\tilde\theta\sqrt{N}}{N\tilde\lambda_1},D_2^N\right)\end{aligned}$$

and $\mu^N(\lfloor y\sqrt{N}\rfloor/\sqrt{N})=\mu N$. Therefore,

$$\begin{aligned}&m^N(y)\\&\quad=\frac{1}{\sqrt{N}}\left[\lambda^N\left(\frac{\lfloor y\sqrt{N}\rfloor}{\sqrt{N}}\right)-\mu^N\left(\frac{\lfloor y\sqrt{N}\rfloor}{\sqrt{N}}\right)\right]\\&\quad=\frac{1}{\sqrt{N}}\left[\frac{\mu N-\mu\sqrt{N}\delta}{g(0,0)}g\left(\frac{\lfloor y\sqrt{N}\rfloor-\tilde\theta\sqrt{N}}{N\tilde\lambda_1},D_2^N\right)-\mu N\right]\\&\quad=-\delta\mu+\frac{\sqrt{N}\mu-\delta\mu}{g(0,0)}\\&\qquad\cdot\left[g\left(\frac{1}{\sqrt{N}}\frac{\lfloor y\sqrt{N}\rfloor/\sqrt{N}-\tilde\theta}{\tilde\lambda_1},D_2^N\right)-g(0,0)\right]\\&\quad\to m^0(y)\quad\text{u.o.c as } N\to\infty.\end{aligned}$$

Similarly,

$$\begin{aligned}&(\sigma^2)^N(y)\\&\quad=\frac{1}{N}\left[\lambda^N\left(\frac{\lfloor y\sqrt{N}\rfloor}{\sqrt{N}}\right)+\mu^N\left(\frac{\lfloor y\sqrt{N}\rfloor}{\sqrt{N}}\right)\right]\\&\quad=\frac{1}{N}\left[\frac{\mu N-\mu\sqrt{N}\delta}{g(0,0)}g\left(\frac{\lfloor y\sqrt{N}\rfloor-\tilde\theta\sqrt{N}}{N\tilde\lambda_1},D_2^N\right)+\mu N\right]\\&\quad\to(\sigma^2)^0(y)\quad\text{u.o.c. as } N\to\infty.\end{aligned}$$

This completes the proof. □

Proof of Proposition 2. The state-space collapse condition of (12) is established by imitating the proof of Propositions 3.1–3.2 in Armony and Maglaras (2004). The main part of that proof analyzes appropriately defined fluid-scaled processes, derives the corresponding limit, and shows that starting from any initial state—not necessarily satisfying (12)—the system reaches the "target state" given in (12) in finite time. By the construction of these fluid-scaled processes, this movement in finite time will appear instantaneous in the natural time scale of the system, and thus asymptotically the system will always be at the appropriate configuration given by (12); this argument is along

the lines of Bramson's state-space collapse result (Bramson 1998).

More specifically, let $A^N(t)$ be the cumulative number of arrivals in both classes up to time t. Note that $A^N(t/\sqrt{N})/\sqrt{N} \to \tilde{\lambda}_1 + \tilde{\lambda}_2 = \mu t$. Hence, the fluid equation derived in the proof of Proposition 3.1 (Armony and Maglaras 2004) for the total queue-length process remains unchanged. Given the definition of $\breve{w}_1^N$ in terms of the total queue-length process, this implies that in fluid scale the limiting λs are constant. As a result, the queue-length equations for each class also remain the same, and the proof follows the exact same steps. Finally, the proof of Proposition 3.2 (Armony and Maglaras 2004) is still valid for the state dependent case, and (13) is immediately established. The last assertion is true because in proving (13) one only uses the first-order term for the arrival rates given by $N\tilde{\lambda}_i$, which is again constant. This completes the proof. □

Proof of Proposition 3. (1) The proof of the asymptotic consistency follows immediately from the continuous mapping theorem, (12), and (13).

(2) From (12) and (13) it follows that

$$\sqrt{N}W_2^N(t) \quad \Rightarrow \quad \widetilde{W}_2(t) = \frac{\widetilde{X}(t)^+ \wedge \tilde{\theta}}{\tilde{\lambda}_2} \leqslant \frac{\tilde{\theta}}{\tilde{\lambda}_2} = \widetilde{D}_2.$$

(3) To establish the asymptotic optimality, consider an arbitrary nonidling nonpreemptive policy π, and denote the threshold policy by π^*. The superscript π will be added to all terms involved to denote their dependence on the policy. If the information announced to customers is $\breve{w}_1^N(\cdot)$ and D_2^N, then the resulting *total* arrival rate into the system given any total number of customers in the system is the same as in the threshold policy. This is due to the fact that the class 1 waiting time estimator $\breve{w}_1^N(\cdot)$ depends on the state of the system only through the total number of customers, and not on the individual queue lengths. Hence, the terms of the infinitesimal drift and variance are the same as $m^N(\cdot)$ and $(\sigma^2)^N(\cdot)$ (defined in the proof of Proposition 1), respectively. Consequently, $X^{N,\pi} \Rightarrow \widetilde{X}$, where $\widetilde{X}$ is the diffusion limit of X^{N,π^*} (as specified in the statement of Proposition 1) given that the threshold policy is used. Let $\widetilde{X}_i^\pi$ be the associated limit queue-length processes that are assumed to exist. Lemma A.2 of Puhalskii and Reiman (2000) is still valid and, as in Proposition 2, $\widetilde{W}_i^\pi(t) = \widetilde{X}_i^\pi(t)/\tilde{\lambda}_i$ for $i = 1, 2$. Because π is asymptotically compliant (i.e., $[\sqrt{N}W_2^{N,\pi}(t) - \widetilde{D}_2]^+ \Rightarrow 0$), it follows that

$$\widetilde{W}_2^\pi(t) \leqslant \widetilde{D}_2 \quad \forall t \geqslant 0 \quad \Leftrightarrow \quad \widetilde{X}_2^\pi(t) \leqslant \tilde{\lambda}_2\widetilde{D}_2 \quad \forall t \geqslant 0.$$

It is easy to show that the proposed threshold policy is pointwise optimal. For every sample path ω, let $\widetilde{X}^+(t,\omega)$ denote the total queue-length trajectory. This is the same for all nonidling policies. Note that for all ω, all $t \geqslant 0$, and for all π such that $\widetilde{X}_2^\pi(t) \leqslant \tilde{\lambda}_2\widetilde{D}_2$,

$$\widetilde{X}_1^{\pi^*}(t,\omega) = \arg\min\{X_1 :\ X_1 + X_2 = \widetilde{X}^+(t,\omega),\ X_2 \leqslant \tilde{\lambda}_2\widetilde{D}_2\} \leqslant \widetilde{X}_1^\pi(t,\omega). \tag{26}$$

Because $\widetilde{W}_1^\pi = \widetilde{X}_1^\pi/\tilde{\lambda}_1$, this completes the proof. Given that (26) holds w.p.1 for all $t \geqslant 0$, it follows that the threshold policy π^* also minimizes (the weaker objective) $\mathbf{E}\widetilde{W}_1$ subject to $\widetilde{W}_2(t) \leqslant \widetilde{D}_2$ for all t. This completes the proof. □

Proof of Corollary 2. Recall that

$$b = -\frac{\kappa}{\gamma}\frac{1}{\tilde{\lambda}_1(0,0)} \quad \text{and} \quad \tilde{\delta} = \delta - \frac{\zeta}{\gamma}\widetilde{D}_2.$$

Note that $\widetilde{X}$ changes behavior in three different regions as follows.

(i) $\widetilde{X} \leqslant 0$: $\widetilde{X}$ behaves like an O-U process. From Halfin and Whitt (1981), $\mathbf{P}(\widetilde{X} \leqslant x \mid \widetilde{X} \leqslant 0) = \Phi(\tilde{\delta} + x)/\Phi(\tilde{\delta})$.

(ii) $\widetilde{X} \in [0, \tilde{\theta}]$: $\widetilde{X}$ behaves like a Brownian motion with drift $-\tilde{\delta}\mu$ and variance 2μ. If $\tilde{\delta} > 0$, the steady-state distribution in $[0, \tilde{\theta}]$ is exponential with rate $\tilde{\delta}$. If $\tilde{\delta} < 0$, then $\tilde{\theta} - \widetilde{X}$ is distributed exponentially with rate $-\tilde{\delta} > 0$; see Browne and Whitt (1995, §18.4.3). If $\tilde{\delta} = 0$, $\widetilde{X}$ is uniform in $[0, \tilde{\theta}]$; see Browne and Whitt (1995, §18.4.2).

(iii) $\widetilde{X} \geqslant \tilde{\theta}$: In this interval $\widetilde{X}$ satisfies the following s.d.e. $d\widetilde{X}_t = -[\tilde{\delta}\mu + b\mu(\widetilde{X}_t - \tilde{\theta})]dt + \sqrt{2\mu}\,dB_t$, where B_t is a standard Brownian motion. Define Y_t such that $b\mu Y = b\mu\widetilde{X} - b\mu\tilde{\theta} + \tilde{\delta}\mu \Rightarrow Y = \widetilde{X} - \tilde{\theta} + \tilde{\delta}/b$. In this case, $dY_t = d\widetilde{X}_t = (-\tilde{\delta}\mu + b\mu\tilde{\theta} - b\mu\widetilde{X}_t)dt + \sqrt{2\mu}\,dB(t) = -b\mu Y_t dt + \sqrt{2\mu}\,dB(t)$. Hence, Y is an O-U process with drift $-b\mu Y(t)$ and variance 2μ. It is well known that the associated steady-state distribution is normal with mean zero and variance $1/b$. It now follows that for $\widetilde{X} \geqslant \tilde{\theta}$, $\widetilde{X} \sim N(-\tilde{\delta}/b + \tilde{\theta}, 1/b)$. That is,

$$\mathbf{P}(\widetilde{X} \geqslant x \mid \widetilde{X} \geqslant \tilde{\theta}) = \frac{\Phi(-\tilde{\delta}/\sqrt{b} - \sqrt{b}(x - \tilde{\theta}))}{\Phi(-\tilde{\delta}/\sqrt{b})}.$$

Hence, when $\tilde{\delta} \neq 0$, the p.d.f. of $\widetilde{X}$ is given by

$$\psi_{\widetilde{X}}(x) = \begin{cases} (1-\alpha_1)\dfrac{\phi(\tilde{\delta} + x)}{\Phi(\tilde{\delta})}, & x < 0, \\ \alpha_2\tilde{\delta}e^{-\tilde{\delta}x}, & x \in [0, \tilde{\theta}], \\ \alpha_3\sqrt{b}\dfrac{\phi(\sqrt{b}(x - \tilde{\theta}) + \tilde{\delta}/\sqrt{b})}{\Phi(-\tilde{\delta}/\sqrt{b})}, & x \geqslant \tilde{\theta}, \end{cases}$$

where $\alpha_1, \alpha_2, \alpha_3 \in [0, 1]$ and satisfy the following continuity and normalization conditions:

$$(1-\alpha_1)\frac{\phi(\tilde{\delta})}{\Phi(\tilde{\delta})} = \alpha_2\tilde{\delta},$$

$$\alpha_2\tilde{\delta}e^{-\tilde{\delta}\tilde{\theta}} = \alpha_3\sqrt{b}\frac{\phi(\tilde{\delta}/\sqrt{b})}{\Phi(-\tilde{\delta}/\sqrt{b})}, \quad \text{and}$$

$$(1-\alpha_1) + \alpha_2(1 - e^{-\tilde{\delta}\tilde{\theta}}) + \alpha_3 = 1.$$

Solving for these constants we get

$$\alpha_2 = \left[\tilde{\delta}\frac{\Phi(\tilde{\delta})}{\phi(\tilde{\delta})} + 1 - e^{-\tilde{\delta}\tilde{\theta}} + \frac{\tilde{\delta}}{\sqrt{b}}e^{-\tilde{\delta}\tilde{\theta}}\frac{\Phi(-\tilde{\delta}/\sqrt{b})}{\phi(\tilde{\delta}/\sqrt{b})}\right]^{-1},$$

$$\alpha_1 = 1 - \alpha_2\tilde{\delta}\frac{\Phi(\tilde{\delta})}{\phi(\tilde{\delta})}, \quad \text{and} \quad \alpha_3 = \frac{\alpha_2\tilde{\delta}}{\sqrt{b}}e^{-\tilde{\delta}\tilde{\theta}}\frac{\Phi(-\tilde{\delta}/\sqrt{b})}{\phi(\tilde{\delta}/\sqrt{b})}.$$

Note that as $\tilde{\theta} \to \infty$, the above result reduces to that of Halfin and Whitt (1981) (presented in §3).

When $\tilde{\delta} = 0$, $\psi_{X^d}(x) = \alpha_2$ for all $x \in [0, \tilde{\theta}]$,

$$2(1-\alpha_1)\phi(0) = \alpha_2, \quad 2\alpha_3\sqrt{b}\phi(0) = \alpha_2, \quad \text{and}$$
$$(1-\alpha_1) + \alpha_2\tilde{\theta} + \alpha_3 = 1,$$

which implies that

$$\alpha_2 = \left[\sqrt{\frac{\pi}{2}} + \tilde{\theta} + \sqrt{\frac{\pi}{2b}}\right]^{-1}, \quad \alpha_1 = 1 - \sqrt{\frac{\pi}{2}}\alpha_2, \quad \text{and} \quad \alpha_3 = \sqrt{\frac{\pi}{2b}}\alpha_2. \tag{27}$$

This completes the proof. □

Proof of Proposition 4. We start by giving a skeleton of the proof. (i) We evaluate $\mathbf{E}(\widetilde{X}^d - \tilde{\theta})^+$. (ii) We define a new static system that replaces the state-dependent drift term $b(\widetilde{X}^d - \tilde{\theta})^+$ (see Corollary 2) by the constant $b\mathbf{E}(\widetilde{X}^d - \tilde{\theta})^+$. This static system is governed by the behavior of the system studied by Halfin and Whitt (1981) (described in §3) where the crucial parameter β is given by $\beta(\mathbf{E}\widetilde{W}_1^d, \widetilde{D}_2)$. (Note that this is similar but not equal to the static system defined in §5.1.1.) Denote by $\widetilde{X}'$ the corresponding total queue-length process and by $\widetilde{W}_1'$ the corresponding waiting time process for class 1 service given by $\widetilde{W}_1' = (\widetilde{X}' - \tilde{\theta})^+/\lambda_1$. (iii) We show that $\mathbf{E}(\widetilde{X}' - \tilde{\theta})^+ \geqslant \mathbf{E}(\widetilde{X}^d - \tilde{\theta})^+$. This implies that $\mathbf{E}\widetilde{W}_1' \geqslant \mathbf{E}\widetilde{W}_1^d$. (iv) Now define the function $h(w) = w - \mathbf{E}[\text{steady-state class 1 waiting time when system announces to all customers } w]$. That is, in our fictitious system the manager announces $\mathbf{E}\widetilde{W}_1^d$, which results in a pair of arrival rates that induces a steady-state expected waiting time for class 1 given by $\mathbf{E}\widetilde{W}_1'$. Now, as was shown in the proof of Proposition 4.1 in Armony and Maglaras (2004), the unique equilibrium of the static system that announces the steady-state expected waiting time information corresponds to the unique solution of the equation $h(w) = 0$. Let us denote this solution as w^* and recognize that $w^* = \mathbf{E}\widetilde{W}_1^s$. From Proposition 4.1 in Armony and Maglaras (2004) we know that $\partial h(w)/\partial w > 0$. Given the result in part (iii) above, $h(\mathbf{E}\widetilde{W}_1^d) < 0$, which by the monotonicity property of $h(\cdot)$ implies that $\mathbf{E}\widetilde{W}_1^d \leqslant w^* = \mathbf{E}\widetilde{W}_1^s$. This completes the proof of the first assertion. The second assertion follows immediately.

(i) Compute $\mathbf{E}(\widetilde{X}^d - \tilde{\theta})^+$. Using the distribution in Corollary 2 along with (27) we find that

$$\begin{aligned}\mathbf{E}(\widetilde{X}^d - \tilde{\theta})^+ &= 2\sqrt{\frac{\pi}{2b}}\alpha_2 \int_{\tilde{\theta}}^{\infty} (x-\tilde{\theta})\sqrt{b}\phi(\sqrt{b}(x-\tilde{\theta}))\,dx \\ &= \frac{\alpha_2\sqrt{2\pi}}{b}\int_0^{\infty} y\phi(y)\,dy \\ &= \frac{1}{b\left[\sqrt{\pi/2} + \tilde{\theta} + \sqrt{\pi/2b}\right]}.\end{aligned}$$

(ii) Define the static system where we replace the linear drift term $b(\widetilde{X}^d - \tilde{\theta})^+$ by its steady-state value $b\mathbf{E}(\widetilde{X}^d - \tilde{\theta})^+$. This corresponds to a system similar to the one analyzed in Halfin and Whitt (1981) described in §3 with $\beta' = \beta(\mathbf{E}\widetilde{W}_1^d, \widetilde{D}_2) = b\mathbf{E}(\widetilde{X}^d - \tilde{\theta})^+$ (recall that $\tilde{\delta} = 0$), for which

$$\mathbf{E}(\widetilde{X}' - \tilde{\theta})^+ = \frac{\alpha(\beta')}{\beta'}e^{-\beta'\tilde{\theta}}.$$

(iii) Hence,

$$\begin{aligned}\frac{\mathbf{E}(\widetilde{X}' - \tilde{\theta})^+}{\mathbf{E}(\widetilde{X}^d - \tilde{\theta})^+} &= \frac{\alpha(\beta')}{\beta'}e^{-\beta'\tilde{\theta}}b\left[\sqrt{\frac{\pi}{2}} + \tilde{\theta} + \sqrt{\frac{\pi}{2b}}\right] \\ &= \alpha(\beta')e^{-\beta'\tilde{\theta}}b\left[\sqrt{\frac{\pi}{2}} + \tilde{\theta} + \sqrt{\frac{\pi}{2b}}\right]^2 \triangleq g(b);\end{aligned}$$

that is, given $\widetilde{D}_2$, this ratio is just a function of $b > 0$ that depends on the choice model. We will show that for all $b > 0$, $g(b) > 1$ by establishing that (a) $\lim_{b\to 0} g(b) > 1$ and (b) $\partial g(b)/\partial b \geqslant 0$. Note that as $b \to 0$, $\beta' \to 0$ and $\alpha(\beta')e^{-\beta'\tilde{\theta}} \to 1$. Hence,

$$\begin{aligned}\lim_{b\to 0} g(b) &= \lim_{b\to 0} b\left[\sqrt{\frac{\pi}{2}} + \tilde{\theta} + \sqrt{\frac{\pi}{2b}}\right]^2 \\ &= \lim_{b\to 0} b\left[\left(\sqrt{\frac{\pi}{2}} + \tilde{\theta}\right)^2 + \frac{\pi}{2b} + 2\sqrt{\frac{\pi}{2b}}\left(\sqrt{\frac{\pi}{2}} + \tilde{\theta}\right)\right] \\ &= \frac{2}{\pi} > 1.\end{aligned}$$

Also,

$$\frac{\partial g(b)}{\partial b} = \frac{\alpha(\beta')}{\beta'^2}e^{-\beta'\theta}\left[1 + b\frac{\partial\beta'}{\partial b}\left(\frac{\partial\alpha(\beta')}{\partial\beta'} - \frac{2}{\beta'} - \tilde{\theta}\right)\right] > 0,$$

where the last inequality follows for $\beta' = -b\mathbf{E}(\widetilde{X}^d - \tilde{\theta})^+$, $\partial\beta'/\partial b < 0$, and (see Armony and Maglaras 2004, Proposition 4.1) $\partial\alpha(\beta')/\partial\beta' < 0$. Hence, $\mathbf{E}(\widetilde{X}^d - \tilde{\theta})^+ < \mathbf{E}(\widetilde{X}' - \tilde{\theta})^+$.

(iv) From the argument outlined above, it now follows that $\mathbf{E}\widetilde{W}_1^d \leqslant \mathbf{E}\widetilde{W}_1^s$.

From $\mathbf{E}\widetilde{W}_1^d \leqslant \mathbf{E}\widetilde{W}_1^s$, it follows that

$$\mathbf{E}\left[-\frac{\kappa}{\gamma}\frac{(\widetilde{X}^d - \tilde{\theta})^+}{\tilde{\lambda}_1(0,0)}\right] = \beta(\mathbf{E}\widetilde{W}_1^d, \widetilde{D}_2) \leqslant \beta(\mathbf{E}\widetilde{W}_1^s, \widetilde{D}_2).$$

Proposition 1 and Proposition 4.3 in Armony and Maglaras (2004) imply that the asymptotic comparison of the aggregate arrival rates amounts to a comparison of the lost throughput as measured by the βs given above. The desired result follows immediately. (Note that the last step does not require the restriction to $\tilde{\delta} = 0$.) This completes the proof. □

Acknowledgments

The authors are grateful to the anonymous associate editor and two referees for their comments that helped improve the exposition of this paper.

References

Anderson, S. P., A. de Palma, J.-F. Thisee. 1996. *Discrete Choice Theory of Product Differentiation*. MIT Press, Cambridge, MA.

Armony, M., C. Maglaras. 2004. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* **52**(2) 271–292.

Billingsley, P. 1968. *Convergence of Probability Measures*. John Wiley and Sons, New York.

Bitran, G., R. Caldentey. 2002. Two-class priority queueing system with state-dependent arrivals. *Queueing Systems* **40** 353–380.

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

Bramson, M. 1998. State space collapse with applications to heavy-traffic limits for multiclass queueing networks. *Queueing Systems* **30** 89–148.

Brandt, A., M. Brandt. 1999. On a two-queue priority system with impatience and its applications to a call center. *Methodology Comput. Appl. Probab.* **1** 191–210.

Browne, S., W. Whitt. 1995. Piecewise-linear diffusion processes. J. H. Dshalalow, ed. *Advances in Queueing: Theory, Methods, and Open Problems*. CRC Press, Boca Raton, FL, 463–480.

Dobson, G., E. Pinker. 2002. The value of sharing lead-time information. Working paper, Simon Business School, University of Rochester, Rochester, NY.

Duenyas, I. 1995. Single facility due date setting with multiple customer classes. *Management Sci.* **41** 608–619.

Duenyas, I., W. J. Hopp. 1995. Quoting customer lead times. *Management Sci.* **41**(1) 43–57.

Gans, N., Y.-P. Zhou. 2003. A call-routing problem with service-level constraints. *Oper. Res.* **51**(2) 255–271.

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

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

Green, L. V., P. J. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.* **37**(1) 84–97.

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

Hassin, R., M. Haviv. 1997. Equilibrium threshold strategies: The case of queues with priorities. *Oper. Res.* **45** 966–973.

Iglehart, D. L. 1965. Limiting diffusion approximations for the many server queue and the repairman problem. *J. Appl. Probab.* **2** 429–441.

Jennings, O., A. Mandelbaum, W. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Sci.* **42**(10) 1383–1394.

Kolesar, P. J., L. V. Green. 1998. Insights on service system design from a normal approximation to Erlang's delay formula. *Prod. Oper. Management* **7** 282–293.

Maglaras, C., J. Van Mieghem. 2004. Admission and sequencing control under delay constraints with applications to GPS and GLQ. *Eur. J. Oper. Res.* To appear.

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Sci.* **49**(8) 1018–1038.

Maglaras, C., A. Zeevi. 2004. Diffusion approximations for a multiclass Markovian service system with "guaranteed" and "best-effort" service levels. *Math. Oper. Res.* To appear.

Mandelbaum, A., G. Pats. 1995. State-dependent queues: Approximations and applications. F. Kelly, R. Williams, eds. *Stochastic Networks*, Vol. 71. *Proc. IMA*, Springer-Verlag, New York, 239–282.

Plambeck, E. L. 2001. Pricing, leadtime quotation and scheduling in a queue with heterogeneous customers. Technical report, Graduate School of Business, Stanford University, Stanford, CA.

Plambeck, E., S. Kumar, J. M. Harrison. 2001. Leadtime constraints in stochastic processing networks under heavy traffic conditions. *Queueing Systems* **39** 23–54.

Puhalskii, A., M. Reiman. 2000. The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* **32**(2) 564–595.

Stone, C. 1963. Limit theorems for random walks, birth and death processes, and diffusion processes. *Illinois J. Math.* **7** 638–660.

Ward, A. R., W. Whitt. 2000. Predicting response times in processor-sharing queues. D. R. McDonald, S. R. E. Turner, eds. *Analysis of Communication Networks: Call Centres, Traffic and Performance*, Vol. 28. *Fields Institute Communications*. American Mathematical Society, Providence, RI.

Whitt, W. 1999a. Improving service by informing customers about anticipated delays. *Management Sci.* **45** 192–207.

Whitt, W. 1999b. Predicting queueing delays. *Management Sci.* **45** 870–888.

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