M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 11, No. 4, Fall 2009, pp. 644–656
ISSN 1523-4614 | EISSN 1526-5498 | 09 | 1104 | 0644
informs
DOI 10.1287/msom.1080.0244

Managing Service Systems with an Offline Waiting Option and Customer Abandonment

Vasiliki Kostami, Amy R. Ward
Information and Operations Management, Marshall School of Business, University of Southern California, Los Angeles, California 90089 {kostami@usc.edu, amyward@usc.edu}

Many service providers offer customers the choice of either waiting in a line or going offline and returning at a dynamically determined future time. The best-known example is the FASTPASS® system at Disneyland. To operate such a system, the service provider must make an upfront decision on how to allocate service capacity between the two lines. Then, during system operation, he must provide estimates of the waiting times for both lines to each arriving customer. The estimation of offline waiting times is complicated by the fact that some offline customers do not return for service at their appointed time. We show that when demand is large and service is fast, for any fixed-capacity allocation decision, the two-dimensional process tracking the number of customers waiting in a line and offline collapses to one dimension, and we characterize the one-dimensional limit process as a reflected diffusion with linear drift. The analytic tractability of this one-dimensional limit process allows us to solve for the capacity allocation that minimizes average cost when there are costs associated with customer abandonments and queueing. We further show that in this limit regime, a simple scheme based on Little's Law to dynamically estimate in line and offline wait times is effective.

Key words: service operations; customer abandonment; customer impatience; reneging; offline waiting; choice models; heavy traffic
History: Received: November 1, 2007; accepted: September 23, 2008. Published online in *Articles in Advance* December 19, 2008.

1. Introduction

An inherent part of the service experience that customers dislike is waiting. In deference to the fact that waiting influences customer evaluation of service (Taylor 1994), service providers aim to minimize wait times. However, it is generally economically infeasible to eliminate waiting. Hence, it is important to manage customers' perceptions of their wait (see, for example, Maister 1985, Katz et al. 1991, Bitran et al. 2008) and to realize that different mechanisms for managing the customer perception of wait time produce different customer reactions (Munichor and Rafaeli 2007).

One factor that influences the psychological cost of waiting is whether the customer physically waits in a line or is offline and free to engage in other activities. In practice, we observe many different implementations of the offline idea. For example, many restaurants give their patrons wireless devices that signal when a table becomes available. In call centers, the idea of giving customers a call-back option was studied by Armony and Maglaras (2004a, b). Cruises and all-inclusive resorts often allow customers to wander while they wait for space to become available in a desired activity. Student health-care clinics may offer noncritical drop-in patients who face a long delay in seeing a doctor or nurse the option of returning later in the day.

Perhaps the best-known real-life example of an offline queue is the FASTPASS® system in Disneyland. For the most popular rides in Disneyland, visitors have a choice. They can either wait in a line or obtain a FASTPASS. The FASTPASS specifies a time at which the visitor can take the ride, making it possible for the customer to visit other parts of the park instead of waiting in a line. The FASTPASS also benefits Disneyland because offline customers may spend money on food or entertainment while they wander around the park. Hence, the offline queue benefits both Disneyland and its customers.

The question that then arises is why Disneyland, or any other service provider, does not offer only offline queueing. One compelling reason to maintain an inline queue in addition to an offline queue is that some customers who join the offline queue become

consumed in other activities and do not return at their appointed time for service. So the inline queue ensures that capacity is not wasted. Also, customers joining the inline queue generally do not leave, and there may be costs other than having idle capacity associated with customers who leave the line. For example, in the amusement park setting, abandoning customers who do not experience certain rides may be foregoing an important element of the park's value proposition and thus be less likely to return (eliminating a future revenue source). Finally, customer preference for an inline or an offline wait may change according to the required amount of waiting associated with each option.

One convenient implementation of offline queueing is having a reservation system. However, for very popular services, reservations tend to fill quickly. This may be acceptable for a restaurant anxious to maintain an image of exclusivity, but it is unacceptable for many service providers. In particular, in an all-inclusive service setting, such as an amusement park, where customers pay a fixed price for access to a number of different attractions, customers expect to be able to visit any attraction of their choosing throughout the course of a day. In fact, Disneyland attempted to implement a reservation system in the mid-1990s but found that early-arriving guests would quickly reserve all available capacity on all the most popular rides. Guests arriving after 11 AM were denied the reservation option (Dickson et al. 2005).

Thus, it is important to investigate service models in which customers can choose between inline and offline queueing at the time of their arrival. In our model, the customers are homogeneous, have linear delay costs that depend on whether the wait is inline or offline, and join the queue that minimizes the cost of waiting. To operate such a system, the service provider must

1. Make an upfront static decision on how to allocate capacity between the inline and the offline queue, and

2. Provide arriving customers with waiting time estimates in real time for both the inline and offline queue.

In some settings, such as a restaurant, where the server is able to communicate with customers, incorrect waiting time estimates can be corrected. However, in other settings, such as Disneyland or any other amusement park, where communication with offline customers is prohibitively difficult, accurately estimating waiting times is essential.

Our objective is to allocate the capacity between the two queues to minimize the average cost, when there are costs associated with customer abandonments and inline queueing and an assumed revenue per customer in the offline queue. The upfront static capacity allocation decision is motivated by the amusement park setting, in which seats on each ride are allocated in predetermined proportions to the inline and the offline queue. We further dynamically derive wait time estimates that depend on that allocation decision using a simple scheme based on Little's Law. The difficulty inherent in making such estimates accurately is complicated by the presence of customers in the offline queue who may abandon, and it is not a priori clear that a simple scheme can work.

The capacity-allocation problem is intractable. However, we can solve it explicitly in a heavy-traffic asymptotic regime in which demand is large and close to the service rate, meaning service times are short. Then, capacity utilization is near 100%. The capacity-allocation problem becomes tractable because there is a reduction in problem dimensionality: the two-dimensional process tracking the number of customers waiting inline and offline collapses to one dimension.

Most amusement parks have hundreds of customers arriving per hour for rides popular that last only minutes. Furthermore, almost every departing train has customers in every available seat. Hence, our heavy traffic analysis is directly applicable to this setting, provided the demand is close to—but does not grossly exceed—the service rate. Eventually, the system departs from the heavy traffic regime, where diffusion approximations are appropriate, and moves to an overloaded regime, where a fluid analysis such as in Whitt (2006) becomes relevant. To understand where our analysis and conclusions break down, we provide numerics showing that the performance of our approximations remains accurate for arrival rates that exceed the service rate by as much as 20% and degrades thereafter.

The remainder of the paper is organized as follows. We first review some relevant literature. In §2, we present our basic model formulation, which is

a single-server queue with general interarrival and service times and both inline and offline waiting. In §3, we solve for the capacity allocation that minimizes average cost as demand becomes large and service fast and demonstrate the accuracy of our solution through simulation. Section 4 that our wait time estimates are correct as demand becomes large and service fast. Section 5 presents concluding remarks.

The proofs of all our results can be found in a technical appendix. We provide the details of how to extend our model and results to a setting that more closely resembles an amusement park ride (specifically, to a setting in which all customers are served in batches at deterministically spaced intervals) in a companion note. (The online supplement to this paper contains the technical appendix and companion note.)

1.1. Literature Review

The service model we analyze is novel because it combines the features of customer abandonment and customer choice. To do this, we have considered a simple scenario in which the customers are homogeneous, the abandonment distribution is given exogeneously, and the delay costs are linear and depend on whether the wait is inline or offline. Previous work that focuses on one of these two features exclusively incorporates heterogeneous customers. Specifically, Mandelbaum and Shimkin (2000) propose a model in which customer abandonment times are determined by each customer optimizing individual utility function, which balances waiting costs against perceived service benefits, and show that the abandonment distribution emerges as an equilibrium point for the model. The models in Armony and Maglaras (2004a, b) do not have customer abandonments but instead focus on how to manage a system in which customers can choose between the equivalent of inline and offline service, where the offline service is guaranteed to be completed within a maximum delay. Both of the aforementioned papers are motivated by call center applications and so are multiserver models.

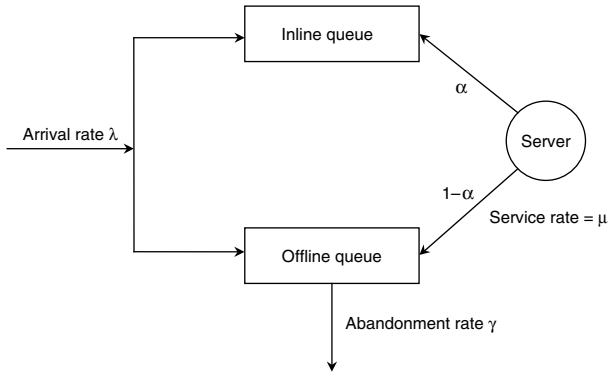
In relation to the queueing literature, our model is a variant of a join-the-shorter queue model. The traditional join-the-shorter queue model that is well studied in the queueing literature has no customer abandonment. For this model, under the assumption of exponential interarrival and service times, the exact solution for the generating function of the stationary distribution of the number of customers in each queue is known, both in the case in which the two service rates are identical (Flatto and McLean 1977) and when they are not (Adan et al. 1991). Results for a join-the-shorter-queue model with general interarrival and service times must rely on an asymptotic analysis, and the heavy traffic analysis is given in Reiman (1984) (along with results on several other models that show state space collapse in a heavy traffic asymptotic regime).

Accurate wait prediction is well studied in the service operations and queueing literature. This is because accurate wait prediction improves customer satisfaction. Whitt (1999) shows how to exploit state information, such as the number of customers ahead of the current customer, to dynamically predict the customer waiting time distribution in a multiserver model that can include customer abandonments.

Our analysis is rougher in the sense that we provide only a point estimate. However, the point estimate is enough for our purposes because in our asymptotic regime the waiting time quotes we provide to customers coincide with the waiting times customers actually experience. In particular, our waiting time quotation policy is asymptotically compliant in the sense of Plambeck et al. (2001). Although the work of Puhalskii (1994) suggests such a result, the presence of customer abandonments complicates the analysis.

Finally, our model considers a single-server system operating in isolation. In an amusement park, as discussed by Ahmadi (1997), there is a larger issue of how to manage capacity and visitor flow throughout the park. It would be interesting to extend Parlakturk and Kumar's (2004) model to investigate how the presence of inline and offline queues and abandonments affects customer routing decisions when there is more than one service station. In general, a complete analytic analysis that incorporates an arbitrary number of service stations with inline and offline queueing and abandonments appears intractable; however, the model formulation in this paper could be used as input into a simulation model such as that developed in Mielke et al. (1998). This would allow for the investigation of further questions of interest from an economics standpoint, such as the one discussed in Oi (1971): Should the amusement

Figure 1 Model



park owner set a two-part tariff in which there is one lump sum admission fee into the park and then separate fees per ride?

2. Model Formulation

We begin our analysis with a single-server system in which each arriving customer chooses between waiting for service in a line or going offline and returning for service at a specified future time point, as shown in Figure 1. The service discipline is head-of-the-line generalized processor sharing. Specifically, when there are customers waiting in both lines, the server processes the customers in the inline queue at rate $\mu\alpha$ and those in the offline queue at rate $\mu(1-\alpha)$, for $\mu > 0$ and $\alpha \in [0, 1]$. We assume each customer in the offline queue may become distracted by other activities while offline and abandon. To model these possible abandonments, we use a nonhomogeneous Poisson process that has rate $\gamma[Q_O(t) - 1]^+$ at time t, for $\gamma > 0$, when the number of customers from the offline queue in the system, including any customer in service, is $Q_O(t)$.

Customers choose which queue to join based on an estimation of the waiting times in the inline and offline queues provided by the server. Note that the server must provide wait time estimations, because customers cannot make their own. This is because customers cannot observe the offline queue themselves, and, in many settings, such as amusement parks, also cannot observe the inline queue. The waiting time estimations we propose are oriented to the situation in which accurate estimation is most important—when demand is large and there is little leftover capacity, meaning there are many customers in both queues. For the inline queue, we estimate the wait time as the expected time to serve all the customers, given that the server is splitting the effort between both queues, so that the wait time estimate at time t is

$$\mathcal{W}_I(t) \equiv \frac{Q_I(t)}{\mu\alpha},$$

where $Q_I(t)$ represents the number of customers from the inline queue in the system, including any customer in service. The parallel waiting time estimation for the offline queue cannot be based on $Q_O(t)$, because the server does not see an offline customer abandon. In particular, the server only realizes the abandonment has occurred after the fact, when the customer fails to return for service at the designated time. So an abandonment cannot be recognized as such by the server until the designated time to receive service has passed. We let $O(t) \geq Q_O(t)$ denote the number of customers in the offline queue recognized by the server, and propose

$$\mathcal{W}_O(t) \equiv \frac{O(t)}{\mu(1-\alpha)}$$

as the waiting time estimation for the offline queue.

In general, we expect that the waiting time quote $\mathcal{W}_O(t)$ will be too high. This is due both to customers who have abandoned the offline queue that the server has not yet seen and to customers who are currently in the offline queue but will eventually abandon and not receive service. However, we will show that such an overestimation is small when demand is large and service is fast. In that case, the arrival and service rates are large compared to the abandonment rate, because the abandonment rate is held fixed. Under these conditions, even though the queue sizes are large, the waiting times are much smaller than the mean abandonment times. Hence, the simple wait time estimation suffices.

We assume customers are homogeneous in their waiting time costs. Let $w_I > 0$ and $w_O > 0$ be the waiting costs per hour for the inline and the offline queues, respectively. A customer arriving at the system at time t minimizes the cost of waiting by joining the inline queue if

$$w_I \mathcal{W}_I(t) \leq w_O \mathcal{W}_O(t)$$

and by joining the offline queue otherwise.

We choose the division of server effort α to minimize infinite horizon average cost. There is a cost, $c > 0$, associated with any customer who abandons who may represent a refund for a service not rendered. There is a holding cost, $h > 0$ per customer, in the inline queue that can be used to penalize the server for the customer's inconvenience. There is a revenue generated, $r > 0$ per customer, in the offline queue that can be used to quantify the value of being free to engage in other activities. We assume $r < c\gamma$ so that the offline queue is costly. In the amusement park setting, the costs c and h represent an expected future revenue loss from the customer being less likely to return at a later date and pay another park entrance fee. The parameter r is actually revenue generated per customer while he wanders around the park, because those customers may purchase food and spend money on entertainment. The total cost after t hours is

$$\mathscr{C}(\alpha, t) \equiv cN\left(\int_0^t \gamma[Q_O(s)-1]^+\,ds\right) + \int_0^t hQ_I(s)\,ds - \int_0^t rQ_O(s)\,ds, \tag{1}$$

where N is a standard Poisson process. We put the α into the notation explicitly to emphasize the dependence of the infinite horizon average cost on it. Let

$$\mathscr{C}(\alpha) \equiv \lim_{t\to\infty} \frac{1}{t}\mathscr{C}(\alpha, t).$$

Our objective is

$$\min_{\alpha\in[0,1]} \mathscr{C}(\alpha). \tag{2}$$

2.1. System Equations

Before considering how to solve the optimization problem (2), we specify the detailed evolution equations for each queue. Let $\{u_i, i \geq 1\}$ be an i.i.d. sequence of nonnegative, mean 1 random variables having finite variance σ_A^2. Let $\{v_i^O, i \geq 1\}$ and $\{v_i^I, i \geq 1\}$ be independent i.i.d. sequences of nonnegative, mean 1 random variables having the same distribution and finite variance σ_S^2. The renewal processes

$$A(t) \equiv \max\left\{i \geq 0:\ \sum_{j=1}^{i} u_j \leq \lambda t\right\},$$

$$S_I(t) \equiv \max\left\{i \geq 0:\ \sum_{j=1}^{i} v_j^I \leq \mu t\right\},$$

$$S_O(t) \equiv \max\left\{i \geq 0:\ \sum_{j=1}^{i} v_j^O \leq \mu t\right\},$$

represent, respectively, the cumulative number of arrivals to the system in $[0, t]$ and the cumulative number of departures from the inline and offline queues after the server has devoted t hours to the queue working at rate μ. Then the evolution equations for Q_I and Q_O are

$$Q_I(t) \equiv \sum_{i=1}^{A(t)} \mathbf{1}\{w_I\mathscr{W}_I(t_i-) \leq w_O\mathscr{W}_O(t_i-)\} - S_I(T_I(t)), \tag{3}$$

$$Q_O(t) \equiv \sum_{i=1}^{A(t)} \mathbf{1}\{w_I\mathscr{W}_I(t_i-) > w_O\mathscr{W}_O(t_i-)\} - N\left(\int_0^t \gamma[Q_O(s)-1]^+\,ds\right) - S_O(T_O(t)), \tag{4}$$

where

$$T_I(t) \equiv \int_0^t \frac{\alpha\mathbf{1}\{Q_I(s) > 0\}}{\alpha + (1-\alpha)\mathbf{1}\{Q_O(s) > 0\}}\,ds, \tag{5}$$

$$T_O(t) \equiv \int_0^t \frac{(1-\alpha)\mathbf{1}\{Q_O(s) > 0\}}{\alpha\mathbf{1}\{Q_I(s) > 0\} + 1 - \alpha}\,ds. \tag{6}$$

Note that $(d/dt)T_I(t)$ and $(d/dt)T_O(t)$ provide the percentage of effort the server allocates to the inline and offline queues, respectively, at time t. When $Q_I(t) > 0$ and $Q_O(t) > 0$, $(d/dt)T_I(t) = \alpha$ and $(d/dt)T_O(t) = (1-\alpha)$. Otherwise, if either $Q_I(t) = 0$ or $Q_O(t) = 0$, but $Q_I(t) + Q_O(t) > 0$, then $(d/dt)T_O(t) = 1$ or $(d/dt)T_I(t) = 1$ accordingly, so that the nonempty queue receives the full server effort.

Define $Q \equiv Q_I + Q_O$ to be the process tracking the total number of customers in the system. The server must work whenever customers are present, and so

$$I(t) \equiv \int_0^t \mathbf{1}\{Q(s) = 0\}\,ds \tag{7}$$

is the cumulative server idle time. Then

$$T_I(t) + T_O(t) + I(t) = t, \tag{8}$$

$$\int_0^\infty Q(t)\,dI(t) = 0. \tag{9}$$

We have made the simplifying assumption that the customers in the offline queue who are served are all present at the service facility when the server is ready to serve them. This is legitimate because we will show that the waiting time estimates we provide are arbitrarily close to the true waiting times experienced by customers in our regime of interest, when demand is large and service is fast (see Theorem 2 in §4). For example, when waiting times are around

one hour, it suffices to ask customers to return to the service facility five minutes before the estimated time at which their service will begin and to assume that serviced customers return at this requested time (see the results of our simulation study in §4).

Note that it is difficult to specify the process O exactly. However, if we let $W_O(t)$ represent the actual waiting time a customer arriving to the offline queue at time t would experience, we can bound the process O as follows:

$$Q_O(t) \le O(t) \le Q_O(t) + N\left(\int_0^t \gamma[Q_O(s)-1]^+ ds\right) - N\left(\int_0^{[t-\sup_{0\le s\le t} W_O(s)]^+} \gamma[Q_O(s)-1]^+ ds\right). \quad (10)$$

The lower bound is obvious. To see the upper bound, realize that all customers who have arrived at the offline queue by time t will have either reached the server or abandoned by time $t + W_O(t)$. Hence, all customers who have arrived at the offline queue by time $[t-\sup_{0\le s\le t} W_O(s)]^+$ will have either reached the server or have abandoned by time

$$\left[t - \sup_{0\le s\le t} W_O(s)\right]^+ + W_O\left(\left[t - \sup_{0\le s\le t} W_O(s)\right]^+\right) \le t.$$

Therefore, the server knows at time t all the customers arriving prior to time $[t-\sup_{0\le s\le t} W_O(s)]^+$ who have abandoned, and the upper bound on O in (10) follows.

3. Revenue Optimization

The capacity-allocation problem (2) cannot be solved with an exact analysis. Even in the case of exponential interarrival and service times, gaining insight from solving the Markov decision problem is difficult. Fortunately, we can solve an approximating problem that becomes accurate in our regime of interest when demand is large and service is fast. In §3.1 we derive the approximating problem, and in §3.2, we solve the approximating problem and verify the accuracy of the solution via simulation.

3.1. Approximating Problem

The key to developing a tractable approximating problem for (2) is to show that the two-dimensional queue-length process can be described by the following one-dimensional diffusion process. Let $\widetilde{X}$ be a Brownian motion having drift $\theta \equiv (\lambda-\mu)/\sqrt{\lambda}$ and variance $\sigma^2 \equiv \sigma_A^2 + \sigma_S^2$. Given $\widetilde{X}$, define the regulated Ornstein-Uhlenbeck process on $[0,\infty)$

$$\widetilde{Q}(t) \equiv \widetilde{X}(t) - \gamma\frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} \cdot \int_0^t \widetilde{Q}(s)\,ds + \tilde{I}(t) \ge 0, \quad (11)$$

for $\tilde{I}$ a nondecreasing process having $\tilde{I}(0)=0$ and $\int_0^\infty \widetilde{Q}(t)\,d\tilde{I}(t)=0$.

Consider a system in which the arrival rate λ becomes large and the service rate is defined as an increasing function of λ. The abandonment rate γ, the server-sharing constant α, and the waiting costs w_I and w_O all remain constant. Our convention is to superscript any process or quantity associated with the system having arrival rate λ by λ.

THEOREM 1. *Consider a system having arrival rate λ and service rate $\mu(\lambda) \equiv \lambda - \sqrt{\lambda}\theta$ for some $\theta \in \Re$.*

(i) *For any $T > 0$, $\sup_{0\le t\le T}(1/\sqrt{\lambda})|(w_I/\alpha)Q_I^\lambda(t) - (w_O/(1-\alpha))Q_O^\lambda(t)| \to 0$, in probability, as $\lambda\to\infty$.*

(ii) *For $(\widetilde{Q},\tilde{I})$ defined by (11) in which $\widetilde{X}$ is a Brownian motion with infinitesimal drift θ and infinitesimal variance σ^2, $(Q^\lambda/\sqrt{\lambda}, I^\lambda/\sqrt{\lambda}) \Rightarrow (\widetilde{Q},\tilde{I})$, as $\lambda\to\infty$.*

The following corollary to Theorem 1 allows us to state a tractable approximating problem for the original capacity-allocation problem (2).

COROLLARY 1. *Consider a system having arrival rate λ and service rate $\mu(\lambda) \equiv \lambda - \sqrt{\lambda}\theta$ for some $\theta \in \Re$. As $\lambda\to\infty$,*

$$\frac{Q_I^\lambda}{\sqrt{\lambda}} \Rightarrow \frac{\alpha w_O}{(1-\alpha)w_I + \alpha w_O}\widetilde{Q},$$

$$\frac{Q_O^\lambda}{\sqrt{\lambda}} \Rightarrow \frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I}\widetilde{Q},$$

$$\frac{N(\int_0^\cdot \gamma[Q_O^\lambda(s)-1]^+ ds)}{\sqrt{\lambda}} \Rightarrow \gamma\frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I}\int_0^\cdot \widetilde{Q}(s)\,ds.$$

Specifically, Corollary 1 suggests that for the total cost up to time t, $\mathscr{C}(\alpha,t)$, defined as in (1),

$$\frac{\mathscr{C}(\alpha,t)}{\sqrt{\lambda}} \approx \left[(c\gamma - r)\frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} + h\frac{\alpha w_O}{\alpha w_O + (1-\alpha)w_I}\right]\int_0^t \widetilde{Q}(s)\,ds.$$

Hence, for large t, letting the random variable $\widetilde{Q}(\infty)$ have the steady-state distribution of the process $\widetilde{Q}$ in (11), and noting that $P(\lim_{t\to\infty} t^{-1}\int_0^t \widetilde{Q}(s)\,ds \to E[\widetilde{Q}(\infty)]) = 1$, it follows from the definition of $\mathscr{C}(\alpha)$ that

$$\frac{1}{\sqrt{\lambda}}\mathscr{C}(\alpha) \approx \left[(c\gamma - r)\frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} + h\frac{\alpha w_O}{\alpha w_O + (1-\alpha)w_I}\right]E[\widetilde{Q}(\infty)].$$

Proposition 18.3 in Browne and Whitt (1995) shows that for ϕ and Φ the density and cumulative distribution functions, respectively, of a standard normal random variable and

$$\kappa \equiv \gamma\frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I},$$

the steady-state mean of the process $\widetilde{Q}$ is

$$E[\widetilde{Q}(\infty)] = \frac{\theta}{\kappa} + \frac{\sigma}{\sqrt{2\kappa}}\frac{\phi((-\theta/\sigma)\sqrt{2/\kappa})}{1-\Phi((-\theta/\sigma)\sqrt{2/\kappa})}. \tag{12}$$

Therefore, defining

$$\widetilde{\mathscr{C}}(\alpha) \equiv \left[(c\gamma - r)\frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} + h\frac{\alpha w_O}{\alpha w_O + (1-\alpha)w_I}\right] \cdot \left(\frac{\theta}{\kappa} + \frac{\sigma}{\sqrt{2\kappa}}\frac{\phi((-\theta/\sigma)\sqrt{2/\kappa})}{1-\Phi((-\theta/\sigma)\sqrt{2/\kappa})}\right),$$

it follows that the problem

$$\min_{\alpha\in[0,1]} \widetilde{\mathscr{C}}(\alpha) \tag{13}$$

approximates the original capacity allocation problem in (2). In particular,

$$\min_{\alpha\in[0,1]} \mathscr{C}(\alpha) \approx \sqrt{\lambda}\min_{\alpha\in[0,1]} \widetilde{\mathscr{C}}(\alpha).$$

Letting $\alpha^\star$ and $\tilde{\alpha}^\star$ denote the respective capacity allocations that result in the minimum cost in (2) and (13), we expect that $\alpha^\star \approx \tilde{\alpha}^\star$.

It is interesting to compare the optimization problem in (13) to the solution for the case when there is no abandonment. In this case, the inline queue is costly, and the offline queue provides revenue, so it is clear that it is optimum to have only an offline queue. We can also see this in the analysis by adapting the objective function in (13) to the case when there is no abandonment, as follows. Similar to the setting in Reiman (1984) (the difference being that his setting has two servers with equal service rates instead of a single server with processor-sharing), Theorem 1 holds, except that the process $\widetilde{Q}$ is a reflected Brownian motion with drift θ and variance σ^2. When $\theta < 0$, the steady-state mean of $\widetilde{Q}$ is $\sigma^2/|\theta|$. (See, for example, Equation (12) in §5.6 in Harrison 1985.) Hence, the objective (13) becomes

$$\min_{\alpha\in[0,1]}\left(-r\frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} + h\frac{\alpha w_O}{\alpha w_O + (1-\alpha)w_I}\right)\frac{\sigma^2}{2|\theta|}.$$

The minimum occurs at $\alpha = 0$, so that having only an offline queue is optimum. In general, the solution to (13) has $\tilde{\alpha}^\star \in [0,1]$. Hence the presence of customer abandonments provides the cost trade-off between inline and offline queueing that makes maintaining both an inline and an offline queue desirable.

3.2. Solution to Approximating Problem

The optimization problem in (13) minimizes a continuous function over a bounded region and so is always solvable numerically. It is easily analytically tractable when $\theta = 0$, and, in §3.2.1, we present the closed form expression for $\tilde{\alpha}^\star$. In §3.2.2, we solve (13) numerically to understand the effect of the cost parameters r, c, h, w_I, and w_O and the capacity parameter θ on $\tilde{\alpha}^\star$. In both subsections, we present simulation results that validate determining the optimum capacity allocation for the original problem (2) by solving the approximating problem (13).

3.2.1. Case: $\theta = 0$. For intuition, we solve (13) in the case that there is exact balance between the arrival and service rates so that $\theta = 0$. Then (13) becomes

$$\min_{\alpha\in[0,1]} f(\alpha), \tag{14}$$

where

$$f(\alpha) \equiv \frac{\sigma}{\sqrt{\pi}\sqrt{\gamma}}\left(\frac{\alpha}{1-\alpha}\frac{w_O}{w_I} + 1\right)^{-1/2}\left(c\gamma - r + h\frac{w_O}{w_I}\frac{\alpha}{1-\alpha}\right).$$

The function f has first derivative

$$f'(\alpha) = \frac{\sigma}{2\sqrt{\pi}\sqrt{\gamma}}\frac{w_O}{w_I}\frac{1}{(1-\alpha)^3}\left(\frac{\alpha}{1-\alpha}\frac{w_O}{w_I} + 1\right)^{-3/2} \cdot \left(\alpha\left(c\gamma - r - 2h + h\frac{w_O}{w_I}\right) + 2h - (c\gamma - r)\right).$$

Table 1 Comparison of Approximated and Simulated Cost to a Simulation

α	Simulated cost ($\mathscr{C}(\alpha)$)	Approximated cost ($\sqrt{\lambda} f(\alpha)$)	Error (%)
0.0	19.006	19.747	3.90
0.1	19.639	19.328	1.59
0.2	19.401	18.923	2.46
0.3	18.733	18.544	1.01
0.4	18.461	18.209	1.36
0.5	18.793	17.952	4.48
0.6	17.264	17.841	3.34
0.7	18.188	18.026	0.89
0.8	18.082	18.923	4.65
0.9	24.031	22.302	7.20

Note. Poisson arrivals with rate 100 per hour, deterministic service with mean 0.01 hours ($\mu = 100$), and parameters $\gamma = 1$, $c = 40$, $h = 10$, $r = 5$, and $w_I = w_O$.

The solution

$$\tilde{\alpha}^\star = \frac{c\gamma - r - 2h}{c\gamma - r - 2h + h(w_O/w_I)}$$

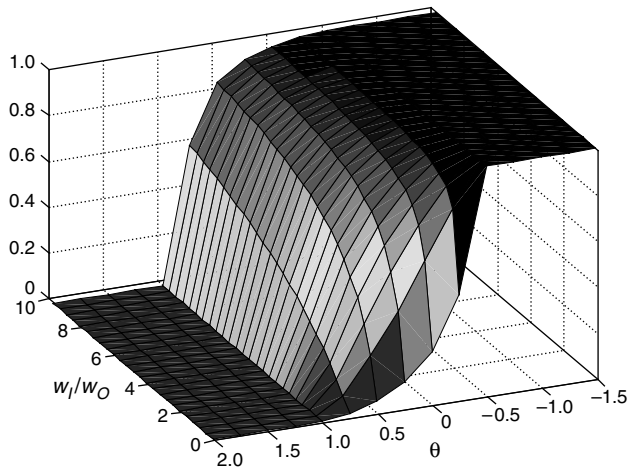
is valid when $2h < (c\gamma - r)$ and has $\tilde{\alpha}^\star \in (0, 1)$. This is because $f'(\tilde{\alpha}^\star) = 0$, $f''(\tilde{\alpha}^\star) > 0$, $f'(0) < 0$, $f(\alpha) \to \infty$ as $\alpha \uparrow 1$, and so $f(\tilde{\alpha}^\star) < f(0)$ and $f(\tilde{\alpha}^\star) < \lim_{\alpha\uparrow 1} f(\alpha)$. Note that when there are no holding costs for the inline queue ($h = 0$), $\tilde{\alpha}^\star = 1$, and so it is optimum to only maintain an inline queue, which matches intuition. Otherwise, in the case that $2h \geq (c\gamma - r)$, it follows that $f'(\alpha) \geq 0$ for all $\alpha \in [0, 1]$. Then the minimum achievable cost occurs at $\tilde{\alpha}^\star = 0$, so having only an offline queue is optimum.

Recall that Corollary 1 suggests that $\mathscr{C}(\alpha) \approx \sqrt{\lambda} f(\alpha)$. Table 1 shows via simulation that the error in this approximation is low (less than 10%) when the system arrival and service rates are 100 or more. (The approximation error decreases as the mean interarrival and service times become shorter, consistent with Corollary 1; however, we do not show these simulation results because of space considerations.) The error sizes in Table 1 are indicative of the error sizes in the approximations suggested by Corollary 1 for both the expected queue lengths and the total number of customer abandonments.

All simulation runs shown in Table 1, and in every table in this paper, are run long enough to generate 10,000,000 arrivals. This ensures that the system has settled into its steady state.

3.2.2. Case: $\theta \neq 0$. In the case that the system is either overloaded or underloaded, we can solve (13) numerically. Figure 2 shows that for any values of w_I and w_O, there exist $\underline{\theta}, \bar{\theta} \in \Re$ such that when $\theta \in (\underline{\theta}, \bar{\theta})$ it is optimal to maintain both an inline and an offline queue ($\tilde{\alpha}^\star \in (0, 1)$). Otherwise, when $\theta \notin (\underline{\theta}, \bar{\theta})$, maintaining only one queue ($\tilde{\alpha}^\star = 0$ or $\tilde{\alpha}^\star = 1$) is optimal.

Figure 2 Solution to (13), $\tilde{\alpha}^\star$, as a Function of θ and w_I/w_O for $c = 40$, $\gamma = 1$, $r = 5$, $h = 7$, and $\sigma = 1$



This is representative of the behavior we find in general, regardless of the specific parameter values of c, γ, r, h, and c. In particular, as θ becomes small, meaning the service capacity is exceeding the arrival rate by more and more, having only an inline queue, $\tilde{\alpha}^\star = 1$, eliminates all abandonment costs and produces a very small inline holding cost, because waiting times are negligible. Otherwise, as θ becomes larger,

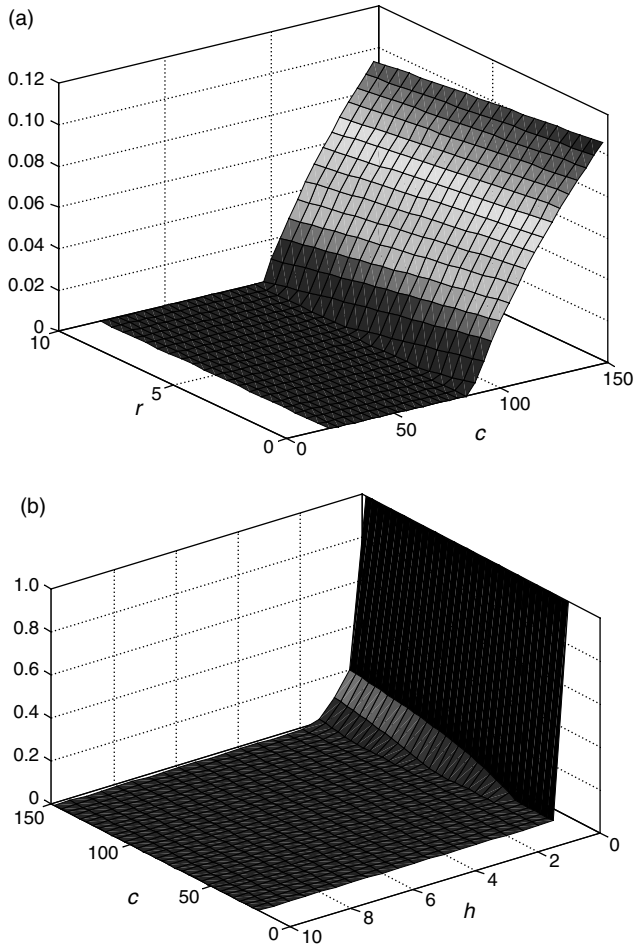
$$\widetilde{\mathscr{C}}(\alpha) \approx \left[(c\gamma - r) \frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} + h \frac{\alpha w_O}{\alpha w_O + (1-\alpha)w_I} \right] \frac{\theta}{\kappa},$$

and the right-hand side is minimized at $\alpha = 0$. Note that θ/κ is the steady-state mean of an unregulated Ornstein-Uhlenbeck process that has the same infinitesimal mean and variance as $\widetilde{Q}$ in (11). The term θ/κ in the preceding display is reflective of the fact that the idleness process in a very heavily loaded system rarely increases; in particular, the process $\widetilde{Q}$ behaves similarly to the unregulated process having $\tilde{I}(t) = 0$ for all $t \geq 0$.

We also observe that Figure 2 is consistent with the solution for $\tilde{\alpha}^\star$ in §3.2.1 when $\theta = 0$. In particular, $\tilde{\alpha}^\star$ increases as the ratio w_I/w_O increases.

When abandonment costs are low, because customers in the offline queue generate revenue, we

Figure 3 **Solution to (13), $\tilde{\alpha}^\star$**



Notes. Top: $\tilde{\alpha}^*$ as a function of c and r for $w_I/w_O = 1$, $\gamma = 1$, $\theta = 2$, $h = 1$, and $\sigma = 1$. Bottom: (b) $\tilde{\alpha}^*$ as a function of c and h for $w_I/w_O = 1$, $\gamma = 1$, $\theta = 2$, $r = 0.5$, and $\sigma = 1$.

expect costs to be minimized by maintaining only an offline queue. Figure 3 confirms this intuition. In particular, when c is low relative to either r or h, $\tilde{\alpha}^\star = 0$, and as c becomes high relative to r or h, $\tilde{\alpha}^\star$ becomes positive and increases.

Corollary 1 suggests that for an unbalanced system ($\theta \neq 0$)

$$\min_{\alpha\in[0,\,1]} \mathscr{C}(\alpha) \approx \sqrt{\lambda} \min_{\alpha\in[0,\,1]} \widetilde{\mathscr{C}}(\alpha),$$

when the arrival rate is within order $\sqrt{\lambda}$ of the service rate. Tables 2 and 3 confirm that this is the case. In particular, our approximation has less than 10% relative error for arrival rates that are as high as 120 per hour when the service rate is 100 per hour. We focus on the case that the arrival rate exceeds the service rate

Table 2 **Comparison of Approximated and Simulated Cost for Overloaded Systems to a Simulation**

λ	$(\lambda-\mu)/\mu$ (%)	α^*	Approximated cost ($\sqrt{\lambda}\widetilde{C}(\tilde{\alpha}^\star)$)	Simulated cost ($\mathscr{C}(\alpha^\star)$)	Error (%)
150	50	0.0	122.47	149.98	18.34
140	40	0.0	101.42	120.08	15.54
130	30	0.0	78.944	90.019	12.30
120	20	0.0	55.076	60.266	8.61
110	10	0.0	32.346	33.529	3.53
109	9	0.1	30.316	31.735	4.47
108	8	0.2	28.288	29.557	4.29
107	7	0.3	26.266	27.184	3.38
106	6	0.4	24.255	25.193	3.72
105	5	0.5	22.269	23.861	6.67
104	4	0.5	20.278	20.839	2.69
103	3	0.6	18.297	19.064	4.03
102	2	0.7	16.374	17.258	5.12
101	1	0.7	14.504	15.108	4.00
100	0	0.8	12.616	13.622	7.39

Note. Poisson arrivals with rate λ per hour, deterministic service with mean 0.01 hours ($\mu = 100$), and parameters $\gamma = 1$, $c = 40$, $h = 5$, $r = 10$, and $w_I = w_O$.

because this is when our approximations are most relevant; otherwise, when the service rate exceeds the arrival rate, the wait times are small (as can be seen in Table 4), so accurate approximation is not as important. Note that to find $\alpha^\star$ we simulated the total cost for various values of α ($\alpha = 0, 0.1, 0.2, \ldots, 0.9$) and chose the minimum cost value. In all cases, the minimum cost value was achieved at exactly the $\tilde{\alpha}^\star$ predicted by our approximation (13). Also note that for the given values of λ and μ, the value of θ is specified as in Corollary 1; i.e., $\theta = \lambda^{-1/2}(\lambda - \mu)$.

We note that Tables 2 and 3 also show that our proposed cost approximation breaks down as the arrival rate increases far past the service rate, by more than 20%. This is not surprising, because the system is moving out of a heavy traffic regime and into an overloaded regime, where a fluid analysis becomes relevant.

4. Waiting Time Quotation

We claimed in §2 that the waiting time estimations we proposed for the inline and offline queues at time t,

$$\mathscr{W}_I(t) = \frac{Q_I(t)}{\mu\alpha} \quad \text{and} \quad \mathscr{W}_O(t) = \frac{O(t)}{\mu(1-\alpha)},$$

were very close to the waiting time a customer joining either queue at time t would experience in our parameter regime of interest, when arrival and service rates

Table 3 Comparison of Approximated Expected Number of Customers in Inline and Offline Queue and Number of Abandonments at Optimal Capacity Allocation to a Simulation

		E[Inline queue length]		E[Offline queue length]		No. of abandonments	
λ	α^*	Approximated	Error (%)	Approximated	Error (%)	Approximated	Error (%)
150	0.0	0	N/A	40.825	18.28	4,082,480	22.51
140	0.0	0	N/A	33.806	15.58	3,380,620	18.23
130	0.0	0	N/A	26.315	12.36	2,631,450	13.99
120	0.0	0	N/A	18.359	8.69	1,835,850	9.64
110	0.0	0	N/A	10.782	3.38	1,078,180	6.16
109	0.1	1.1024	4.87	9.9218	4.70	992,176	3.94
108	0.2	2.2631	0.73	9.0523	4.67	905,231	3.10
107	0.3	3.5021	0.97	8.1715	3.62	817,154	3.19
106	0.4	4.8510	5.73	7.2765	3.49	727,649	2.30
105	0.5	6.3624	8.83	6.3625	2.12	636,255	0.56
104	0.5	5.7936	5.01	5.7936	2.26	579,364	1.62
103	0.6	7.3186	6.62	4.8791	3.33	487,907	0.45
102	0.7	9.1696	8.94	3.9298	3.86	392,983	1.70
101	0.7	8.1221	2.39	3.4809	4.63	348,089	3.66
100	0.8	10.093	13.69	2.5231	7.51	252,313	3.91

Note. Poisson arrivals with rate λ per hour, deterministic service with mean 0.01 ($\mu = 100$), and parameters $\gamma = 1$, $c = 40$, $h = 5$, $r = 10$, and $w_I = w_O$.

are close and large compared to the abandonment rate. This is not surprising for the inline queue. However, this is not obvious for the offline queue, because some customers in the offline queue may abandon, and the process O bounded in (10) includes customers who have already abandoned the offline queue but of whom the server is not yet aware.

Our next theorem shows that $\mathcal{W}_I(t)$ is very close to the actual waiting time a customer joining the inline queue at time t would experience, which we denote by $W_I(t)$, and that $\mathcal{W}_O(t)$ is very close to the actual waiting time a customer joining the offline queue at time t would experience, which we denote by $W_O(t)$. As in Theorem 1, we consider a system in which the arrival rate λ becomes large and the service rate is defined as an increasing function of λ, and we superscript any process or quantity associated with the system having arrival rate λ by λ.

Table 4 Comparison of Actual and Quoted Wait Times for Inline and Offline Queues to a Simulation

	Inline queue		Offline queue		
λ	Avg. wait	Avg. absolute difference	Avg. wait	Avg. absolute difference	P(abandon)
150	61.57	0.0133	41.65	18.5623	0.50
140	49.38	0.0223	35.30	12.8790	0.44
130	37.37	0.0216	28.30	8.0390	0.38
120	25.23	0.0197	20.27	4.0955	0.29
110	13.50	0.0170	11.39	1.4332	0.17
100	5.96	0.0401	4.99	0.3477	0.08

Note. Poisson arrivals with rate λ per hour, deterministic service with mean 0.01 hours ($\mu = 100$), and parameters $\gamma = 1$, $\alpha = 0.5$, and $w_I = w_O = 1$.

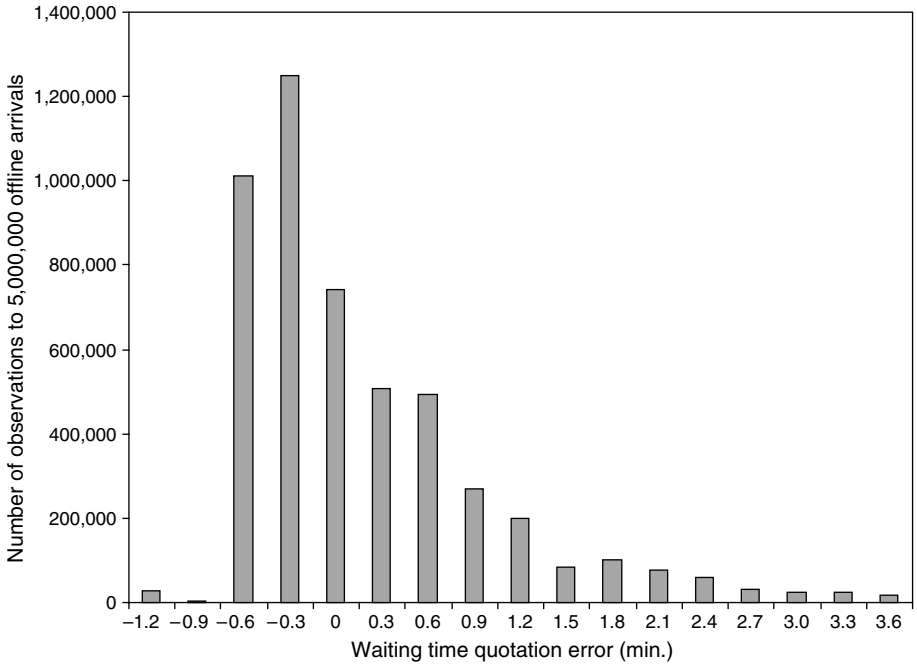
THEOREM 2. *Consider a system having arrival rate λ and service rate $\mu(\lambda) \equiv \lambda - \sqrt{\lambda}\theta$ for some $\theta \in \Re$. For any $T > 0$, as $\lambda \to \infty$*

$$\sup_{0 \le t \le T} \sqrt{\lambda} |W_I^\lambda(t) - \mathcal{W}_I^\lambda(t)| \to 0 \quad \textit{and}$$

$$\sup_{0 \le t \le T} \sqrt{\lambda} |W_O^\lambda(t) - \mathcal{W}_O^\lambda(t)| \to 0, \ \textit{in probability}.$$

Theorem 2 shows that our waiting time quotations are very accurate when the arrival rate λ is large and within $\sqrt{\lambda}$ of the service rate. As in §3.2.2 when considering the accuracy of our proposed cost function approximation, we would also like to understand how our proposed waiting time quotations perform as the arrival rate increases past the service rate. To do this, in Table 4 we simulate a system having fixed capacity allocation and for every arriving customer, we record the actual wait time and the wait time quote in minutes. We then report the average actual inline and offline wait times and the average absolute difference between the actual and quoted wait times in minutes. We expect that the inline wait time quotes will be very accurate, even outside the parameter

Figure 4 Histogram of Difference Between Waiting Time Quotation and Actual Waiting Time



Note. $\alpha = 0.5$ in a simulation having Poisson arrivals with rate 100 per hour, deterministic service with mean 0.01, and parameters $\gamma = 0.01$, and $w_l = w_0 = 1$.

regime stated in Theorem 2. Because the service times are deterministic, eliminating an inherent system variability, the only reasons the inline wait time quotes should not match actual wait times would be because of the residual service times and a possibly empty offline queue. We include them in Table 4 as a benchmark for comparison purposes.

We conclude that the offline wait time quotes are very accurate for parameters that satisfy the conditions of Theorem 2. (When the arrival rate is less than the service rate, the wait times in both queues are very small, so accurate wait time quotation is not so important.) It is also true that the accuracy of the offline wait quotes decreases monotonically as the arrival rate increases past the service rate. This is because the percentage of customers abandoning the offline queue increases monotonically as the arrival rate increases past the service rate.

Recall that the system evolution equations in (3)–(9) make the simplifying assumption that the customers in the offline queue who are served are all present at the service facility when the server is ready to serve them. How can we ensure that this is indeed the case? Theorem 2 suggests that for a system in which the arrival and service rates are close and large compared to the abandonment rate, we can simply ask customers to return a little before their estimated service time. Figure 4 quantifies the meaning of "a little" for one particular example in which the arrival and service rates are 100 customers per hour, the average offline wait time is 47.22 minutes, and the probability a customer abandons is 0.82%, meaning $\gamma = 0.01$. (Note that we have changed the abandonment rate from that in the previous paragraph so that the average wait time in the offline queue will be more than 5 minutes.) The largest wait time quotation error we see over 5,000,000 customer arrivals to the offline queue is 10 minutes; therefore, if we have customers return to the service facility 10 minutes before their estimated service time, we would ensure that all served customers are present at the service facility when the server is ready to serve them.

In general, the amount of time before the estimated offline wait time that customers must return to the service facility to ensure their presence when it is desired varies according to the system parameters and the average offline waiting time. Suppose we ask a customer who chooses to join the offline queue at time t to

Table 5 Value of ϵ Such That $\mathcal{W}_O(t) - \epsilon < W_O(t)$ for 95%, 98%, and 99% of Customers Joining Offline Queue

	ϵ in min to achieve (%)			Simulated avg. offline wait
α	95	98	99	(in min)
0.0	0.51	0.10	1.24	34.72
0.1	0.76	1.08	1.41	35.87
0.2	0.62	1.40	1.79	38.92
0.3	1.06	1.44	1.81	39.51
0.4	1.34	1.76	2.18	43.11
0.5	1.67	2.15	3.11	47.22
0.6	2.13	3.26	3.83	56.26
0.7	2.89	3.65	5.16	60.56
0.8	5.34	6.61	7.89	80.91
0.9	7.81	12.41	17.01	105.73

Note. Simulation has Poisson arrivals with rate 100 per hour, deterministic service with mean 0.01 hours ($\mu = 100$), abandonment rate $\gamma = 0.01$, and $w_I = w_O = 1$.

return to the service facility at time $\mathcal{W}_O(t) - \epsilon$. Table 5 shows what the value of ϵ must be to ensure that 95%, 98%, and 99% of the customers choosing to join the offline queue are present at the service facility when desired. For example, in a system having arrival and service rates of 100 customers per hour, abandonment rate $\gamma = 0.01$, and $\alpha = 0$ (so there is only an offline queue), asking customers to return 1.24 minutes before their estimated service time ensures 99% of the customers are present when required. This 1.24 minutes is a negligible amount of "padding" when the average offline wait time is 34.72 minutes. Of course, the system evolution equations in (3)–(9) are not exactly modeling what happens for the small percentage of customers for which we grossly err, i.e., for those for which $\mathcal{W}_O(t) - \epsilon > W_O(t)$. Are these customers effectively abandoned? Are they absorbed into the offline queue when they appear? In either case, when their percentage is small enough, their effect on the overall system behavior is negligible, so our model is representative of the overall system behavior.

5. Conclusion

In this paper, we analyze a single-server system with two waiting modes: inline and offline. Customers have linear delay costs and pick the mode with the smaller delay cost based on their waiting time quote. The customers who join the offline queue may abandon. We show that when demand is large and service is fast, the two-dimensional process tracking the number of customers waiting in line and offline can be described by a one-dimensional reflected diffusion with linear drift. The analytic tractability of this limit process allows us to provide an approximation of the capacity allocation that minimizes the average cost. Moreover, we can accurately predict the waiting time of any new arrival using a simple scheme based on Little's Law, despite the abandonments that may occur in the offline queue. We demonstrate the accuracy of our approximations via simulation.

We end by noting that our results continue to hold in a setting that more closely models an amusement park ride, in which customers are served in batches at discrete time points. For the details of this setting, we refer the interested reader to our companion note.

Electronic Companion

An electronic companion to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

References

Adan, I. J., J. Wessels, W. H. M. Zijm. 1991. Analysis of the asymmetric shortest queue problem. *Queueing Systems* **8** 1–58.

Ahmadi, R. H. 1997. Managing capacity and flow at theme parks. *Oper. Res.* **45**(1) 1–13.

Armony, M., C. Maglaras. 2004a. Contact centers with a call-back option and real-time delay information. *Oper. Res.* **52**(4) 527–545.

Armony, M., C. Maglaras. 2004b. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* **52**(4) 271–292.

Bitran, G. R., J. C. Ferrer, P. R. Oliveira. 2008. Managing customers experiences: Perspectives on the temporal aspects of service encounters. *Manufacuring Service Oper. Management* **10**(1) 61–83.

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.

Dickson, D., R. C. Ford, B. Laval. 2005. Managing real and virtual wait in hospitality and service organizations. *Cornell Hotel Restaurant Admin. Quart.* **46** 52–68.

Flatto, L., H. P. McLean. 1977. Two queue in parallel. *Comm. Pure Appl. Math.* **30** 255–263.

Harrison, J. M. 1985. *Brownian Motion and Stochastic Flow Systems*. Krieger, Malabar, FL.

Katz, K., B. Larson, R. Larson. 1991. Prescription for the waiting in line blues: Entertain, enlighten and engage. *Sloan Management Rev.* **32**(2) 44–53.

Maister, D. 1985. The psychology of waiting in lines. J. A. Czepiel, M. Solomon, C. S. Surprenant, eds. *The Service Encounter*. Lexington Books, Lexington, MA, 113–123.

Mandelbaum, A., N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems* **36** 141–173.

Mielke, R., A. Zahralddin, D. Padam, T. Mastaglio. 1998. Simulation applied to theme park management. D. J. Medeiros, E. F. Watson, J. S. Carson, M. S. Manivannan, eds. *Proc. 1998 Winter Simulation Conf.*, 1199–1203.

Munichor, N., A. Rafaeli. 2007. Number of apologies? Customer reactions to telephone waiting time fillers. *J. Applied Psych.* **92**(2) 511–518.

Oi, W. I. 1971. A Disneyland dilemma: Two-part tariffs for a Mickey Mouse monopoly. *Quart. J. Econom.* **85**(1) 77–96.

Parlakturk, A., S. Kumar. 2004. Self-interested routing in queueing networks. *Management Sci.* **50**(7) 949–967.

Plambeck, E., S. Kumar, J. M. Harrison. 2001. A multiclass queue in heavy traffic with throughput time constraints; asymptotically optimal dynamic controls. *Queueing Systems* **39** 23–54.

Puhalskii, A. 1994. On the invariance principle for the first passage time. *Math. Oper. Res.* **19** 946–954.

Reiman, M. I. 1984. Some diffusion approximations with state space collapse. F. Bacceli, G. Fayolle, eds. *Modelling and Performance Evaluation Methodology*. Springer-Verlag, New York, 209–240.

Taylor, S. 1994. Waiting for service: The relationship between delays and evaluations of service. *J. Marketing* **58**(2) 56–69.

Whitt, W. 1999. Improving service by informing customers about anticipated delays. *Management Sci.* **45**(2) 192–207.

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.