

Managing Service Systems with an Offline Waiting Option and Customer Abandonment

Vasiliki Kostami • Amy R. Ward

*Information and Operations Management,
Marshall School of Business, USC,
3670 Trousdale Parkway,
Los Angeles, CA, 90089-0809, USA
kostami@usc.edu • amyward@usc.edu*

July 23, 2008

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 first make an upfront decision on how to allocate service capacity between the two lines. Then, during system operation, he must dynamically provide estimates of the waiting times at 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 inline and offline collapses to one dimension, and 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 inline and offline wait times is effective.

1. Introduction

An inherent part of the service experience disliked by customers 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 entirely. Hence it is important to manage customers' perceptions of their wait (see, for example, Maister 1985, Katz et al. 1991, Bitran et al 2007), and 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)(2004b). Cruises and all-inclusive resorts often allow customers to wander while waiting for space to become available in a desired activity. Student healthcare clinics may offer non-critical drop-in patients that face a long delay to see 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 wandering 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 that join the offline queue become consumed in other activities, and do not return at their appointed time for service. Hence the inline queue ensures capacity is not wasted. Also, customers joining the inline queue generally do not abandon, and there may be costs other than having idle capacity associated with abandoning customers. For example, in the amusement park setting, abandoning customers that do not experience certain rides may be foregoing an important element of the parks 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 services that are very popular, reservations tend to fill quickly. This may be an acceptable situation for a restaurant anxious to maintain an image of exclusivity, but it is not an acceptable situation 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 1990's but found that early-arriving guests would quickly book all available reservation capacity on all their most popular rides. Guests arriving after

11 am were denied the reservation option (Dickson et al. 2005).

Hence it is of importance 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 if the wait is inline or offline, and join the queue that minimizes the cost of waiting. In order 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 in order 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 in each ride are allocated in pre-determined 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 that 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 the capacity allocation problem 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 is 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 popular amusement park rides have hundreds of customers arriving per hour to ride the ride, and 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 not grossly exceeding, 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 (see Tables 2 and 3) that show the performance of our approximations remain 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 Section 2, we present our basic model formulation, which is a single-server queue with general inter-arrival and service times, and both inline and offline waiting. In Section 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 validates that our wait time estimates are correct as demand becomes large and service fast. Section 5 shows that all our results remain valid when customers are served in batches, at deterministically spaced intervals, as is true for an amusement park ride. We end by making some concluding remarks in Section 6.

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 if 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 his individual utility function, that 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) (2004b) do not have customer abandonments, and 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 multi-server 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 inter-arrival 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 inter-arrival

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).

The issue of accurate wait prediction is well-studied in both 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 only focus on the conditional mean. However, the mean 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 in Ahmadi (1997), there is a larger issue of how to manage capacity and visitor flow throughout the park. It would be interesting to extend the model of Parlakturk and Kumar (2004) 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 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 dynamically 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 possi-

ble abandonments, we use a non-homogeneous 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 $Q_O(t)$.

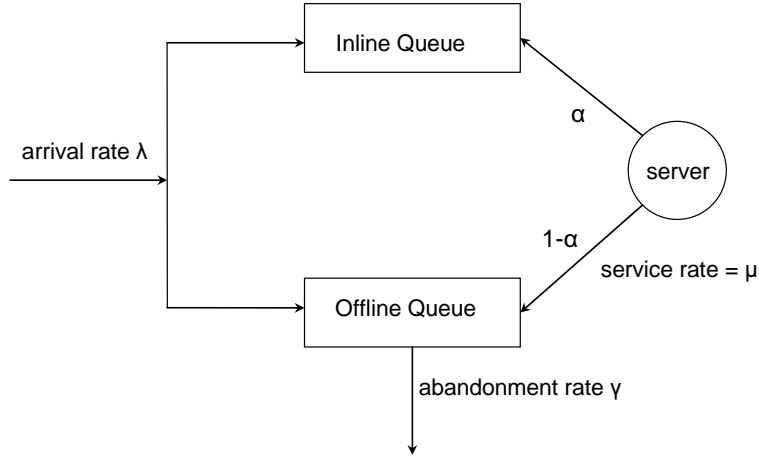


Figure 1: The model

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 due to the fact that 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 his 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 his service at his 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 as 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 that have abandoned the offline queue that the server has not yet seen, and customers that 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. Then, 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, and let $w_I > 0$ and $w_O > 0$ be the waiting costs per hour for the inline and the offline queues respectively. Then, a customer arriving to the system at time t minimizes his 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 would like to choose the division of server effort α in order to minimize infinite horizon average cost. There is a cost $c > 0$ associated with any customer that abandons that 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 due to the customer being less likely to return at a later date and pay another park entrance fee. The parameter r is actually a revenue generated per customer while wandering around the park, because those customers may purchase food and spend money on entertainment. The total cost after t hours is

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

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

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

Our objective is

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

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 non-negative, 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 non-negative, mean 1 random variables having the same distribution and finite variance σ_S^2 . The renewal processes

$$\begin{aligned} A(t) &\equiv \max\{i \geq 0 : \sum_{j=1}^i u_j \leq \lambda t\} \\ S_I(t) &\equiv \max\{i \geq 0 : \sum_{j=1}^i v_j^I \leq \mu t\} \\ S_O(t) &\equiv \max\{i \geq 0 : \sum_{j=1}^i v_j^O \leq \mu t\} \end{aligned}$$

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

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

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

where

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

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

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

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

is the cumulative server idletime. Then,

$$T_I(t) + T_O(t) + I(t) = t \quad (8)$$

$$\int_0^\infty Q(t) dI(t) = 0. \quad (9)$$

We have made the simplifying assumption that the customers in the offline queue that 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 Section 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 Table 5 in Section 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

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

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

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

Therefore, the server knows at time t of all the customers arriving prior to time $[t - \sup_{0 \leq s \leq t} W_O(s)]^+$ that 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 inter-arrival 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 Subsection 3.1 we derive the approximating problem, and in Subsection 3.2, we solve the approximating problem and verify the accuracy of the solution via simulation.

3.1 The 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 \tilde{X} be a Brownian motion having drift $\theta \equiv \frac{\lambda - \mu}{\sqrt{\lambda}}$ and variance $\sigma^2 \equiv \sigma_A^2 + \sigma_S^2$. Given \tilde{X} , define the regulated Ornstein-Uhlenbeck process on $[0, \infty)$

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

for \tilde{I} a non-decreasing process having $\tilde{I}(0) = 0$ and $\int_0^\infty \tilde{Q}(t) d\tilde{I}(t) = 0$.

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 $\theta \in \mathfrak{R}$.

- (i) For any $T > 0$, $\sup_{0 \leq t \leq T} \frac{1}{\sqrt{\lambda}} \left| \frac{w_I}{\alpha} Q_I^\lambda(t) - \frac{w_O}{1-\alpha} Q_O^\lambda(t) \right| \rightarrow 0$, in probability, as $\lambda \rightarrow \infty$.
- (ii) For (\tilde{Q}, \tilde{I}) defined by (11) in which \tilde{X} is a Brownian motion with infinitesimal drift θ and infinitesimal variance σ^2 , $\left(\frac{Q^\lambda}{\sqrt{\lambda}}, \frac{I^\lambda}{\sqrt{\lambda}} \right) \Rightarrow (\tilde{Q}, \tilde{I})$, as $\lambda \rightarrow \infty$.

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 $\theta \in \mathfrak{R}$. As $\lambda \rightarrow \infty$,

$$\begin{aligned} \frac{Q_I^\lambda}{\sqrt{\lambda}} &\Rightarrow \frac{\alpha w_O}{(1-\alpha)w_I + \alpha w_O} \tilde{Q} \\ \frac{Q_O^\lambda}{\sqrt{\lambda}} &\Rightarrow \frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} \tilde{Q} \\ \frac{N \left(\int_0^\cdot \gamma [Q_O^\lambda(s) - 1]^+ ds \right)}{\sqrt{\lambda}} &\Rightarrow \gamma \frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} \int_0^\cdot \tilde{Q}(s) ds. \end{aligned}$$

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

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

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

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

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 \tilde{Q} is

$$E[\tilde{Q}(\infty)] = \frac{\theta}{\kappa} + \frac{\sigma}{\sqrt{2\kappa}} \frac{\phi\left(\frac{-\theta}{\sigma}\sqrt{\frac{2}{\kappa}}\right)}{1 - \Phi\left(\frac{-\theta}{\sigma}\sqrt{\frac{2}{\kappa}}\right)}. \quad (12)$$

Therefore, defining

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

it follows that the problem

$$\min_{\alpha \in [0,1]} \tilde{\mathcal{C}}(\alpha) \quad (13)$$

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

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

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

$$\alpha^* \approx \tilde{\alpha}^*.$$

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 Section 5 in Reiman (1984) (the difference being that his setting has 2 servers with equal service rates instead of a single server with processor-sharing), Theorem 1 holds except that the process \tilde{Q} is a reflected Brownian motion with drift θ and variance σ^2 . When $\theta < 0$, the steady-state mean of \tilde{Q} is $\sigma^2/|\theta|$. (See, for example, equation (12) in Section 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 $\tilde{\alpha}^* \in [0, 1]$. 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 The Solution to the 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 Subsection 3.2.1, we present the closed form expression for $\tilde{\alpha}^*$. In Subsection 3.2.2, we solve (13) numerically in order to understand the effect of the cost parameters r , c , h , w_I , and w_O , and the capacity parameter θ , on $\tilde{\alpha}^*$. 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 The case that $\theta = 0$.

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

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

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

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

The solution

$$\tilde{\alpha}^* = \frac{c\gamma - r - 2h}{c\gamma - r - 2h + h \frac{w_O}{w_I}}$$

is valid when

$$2h < (c\gamma - r),$$

and has $\tilde{\alpha}^* \in (0, 1)$. This is because $f'(\tilde{\alpha}^*) = 0$, $f''(\tilde{\alpha}^*) > 0$, $f'(0) < 0$, $f(\alpha) \rightarrow \infty$ as $\alpha \uparrow 1$, and so $f(\tilde{\alpha}^*) < f(0)$ and $f(\tilde{\alpha}^*) < \lim_{\alpha \uparrow 1} f(\alpha)$. Note that when there are no holding costs for the inline queue ($h = 0$), $\tilde{\alpha}^* = 1$, 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 $\tilde{\alpha}^* = 0$, and so having only an offline queue is optimum.

α	Simulated Cost ($\mathcal{C}(\alpha)$)	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%

Table 1: A comparison of the approximated and the simulated cost to a simulation having 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 $w_I = w_O$.

Recall that Corollary 1 suggests that

$$\mathcal{C}(\alpha) \approx \sqrt{\lambda}f(\alpha).$$

Table 1 shows via simulation that the error in this approximation is low (less than 10%) when the system arrival and service rates are one hundred or more. (The approximation error decreases as the mean inter-arrival and service times become shorter, consistent with Corollary 1; however, we do not show these simulation results due to 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 The case that $\theta \neq 0$.

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 $\underline{\theta}, \bar{\theta} \in \mathfrak{R}$ such that when $\theta \in (\underline{\theta}, \bar{\theta})$ it is optimal to maintain both an inline and an offline queue ($\tilde{\alpha}^* \in (0, 1)$). Otherwise, when $\theta \notin (\underline{\theta}, \bar{\theta})$, maintaining only one queue ($\tilde{\alpha}^* = 0$ or $\tilde{\alpha}^* = 1$) is optimal.

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, $\tilde{\alpha}^* = 1$, eliminates all abandonment costs and produces a very small inline holding cost, because waiting

times are negligible. Otherwise, as θ becomes larger,

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

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 \tilde{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 \tilde{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 $\tilde{\alpha}^*$ in Subsection 3.2.1 when $\theta = 0$. In particular, $\tilde{\alpha}^*$ increases as the ratio w_I/w_O increases.

When abandonment costs are low, because customers in the offline queue generate revenue, we 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 , $\tilde{\alpha}^* = 0$, and as c becomes high relative to r or h , $\tilde{\alpha}^*$ becomes positive and increases.

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

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

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 because this is when our approximations are

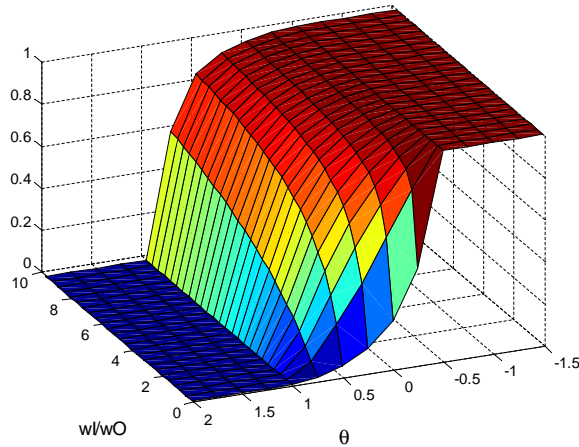
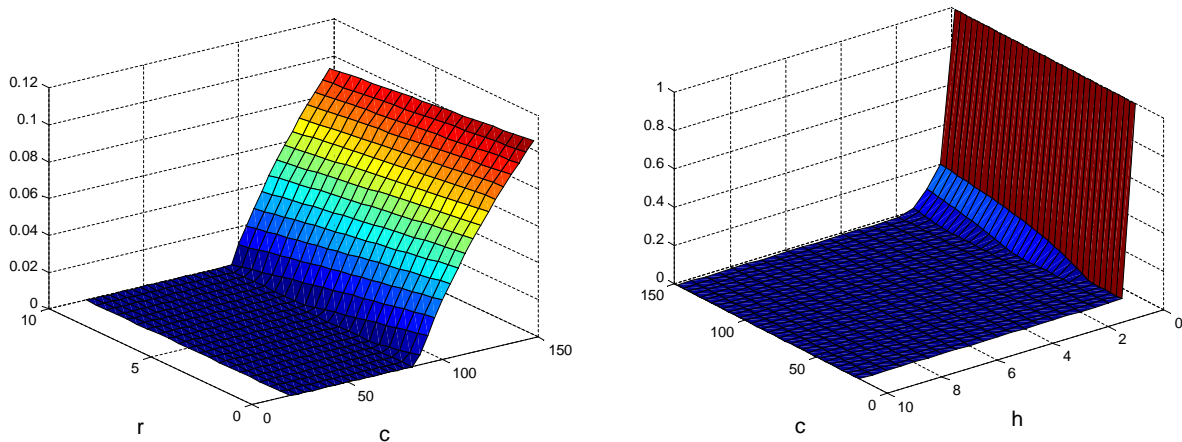


Figure 2: The solution to (13), $\tilde{\alpha}^*$, as a function of θ and $\frac{w_I}{w_O}$ for $c = 40, \gamma = 1, r = 5, h = 7$, and $\sigma = 1$.



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

Figure 3: The solution to (13), $\tilde{\alpha}^*$.

most relevant; otherwise, when the service rate exceeds the arrival rate, the wait times are small (as can be seen in Table 4), and so accurate approximation is not as important. Note that to find α^* we simulated the total cost for various values of α ($\alpha = 0, 0.1, 0.2, \dots, 0.9$), and chose the minimum cost value. In all cases, the minimum cost value was achieved at exactly the $\tilde{\alpha}^*$ 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 Section 2 that the waiting time estimations we proposed for the inline and offline queues at time t ,

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

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 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 (1) some customers in the offline queue

λ	$\frac{\lambda-\mu}{\mu}$	α^*	Approximated Cost ($\sqrt{\lambda}\tilde{C}(\tilde{\alpha}^*)$)	Simulated Cost ($\mathcal{C}(\alpha^*)$)	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%

Table 2: A comparison of the approximated and the simulated cost for overloaded systems to a simulation having 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 $w_I = w_O$.

λ	α^*	E[Inline Queue-length]		E[Offline Queue-length]		# of Abandonments	
		Approximated	Error	Approximated	Error	Approximated	Error
150	0.0	0	N/A	40.825	18.28%	4082480	22.51%
140	0.0	0	N/A	33.806	15.58%	3380620	18.23%
130	0.0	0	N/A	26.315	12.36%	2631450	13.99%
120	0.0	0	N/A	18.359	8.69%	1835850	9.64%
110	0.0	0	N/A	10.782	3.38%	1078180	6.16%
109	0.1	1.1024	4.87%	9.9218	4.70%	992176	3.94%
108	0.2	2.2631	0.73%	9.0523	4.67%	905231	3.10%
107	0.3	3.5021	0.97%	8.1715	3.62%	817154	3.19%
106	0.4	4.8510	5.73%	7.2765	3.49%	727649	2.30%
105	0.5	6.3624	8.83%	6.3625	2.12%	636255	0.56%
104	0.5	5.7936	5.01%	5.7936	2.26%	579364	1.62%
103	0.6	7.3186	6.62%	4.8791	3.33%	487907	0.45%
102	0.7	9.1696	8.94%	3.9298	3.86%	392983	1.70%
101	0.7	8.1221	2.39%	3.4809	4.63%	348089	3.66%
100	0.8	10.093	13.69%	2.5231	7.51%	252313	3.91%

Table 3: A comparison of the approximated expected number of customers in the inline and offline queue and the number of abandonments at the optimal capacity allocation to a simulation having 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 $w_I = w_O$.

may abandon, and (2) the process O bounded in (10) includes customers that have already abandoned the offline queue 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 .

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

$$\sup_{0 \leq t \leq T} \sqrt{\lambda} |W_I^\lambda(t) - \mathcal{W}_I^\lambda(t)| \rightarrow 0 \text{ and } \sup_{0 \leq t \leq T} \sqrt{\lambda} |W_O^\lambda(t) - \mathcal{W}_O^\lambda(t)| \rightarrow 0, \text{ in probability.}$$

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

λ	Inline Queue		Offline Queue		P(abandon)
	Avg. Wait	Avg. Abs. Difference	Avg. Wait	Avg. Abs. Difference	
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

Table 4: A comparison of the actual and quoted wait times for the inline and offline queues to a simulation having Poisson arrivals with rate λ per hour, deterministic service with mean 0.01 hours ($\mu = 100$), and parameters $\gamma = 1$, $\alpha = 0.5$ and $w_I = w_O = 1$.

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, and 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 due to the fact that the percentage of customers abandoning the offline queue increases monotonically as the arrival rate increases past the service rate.

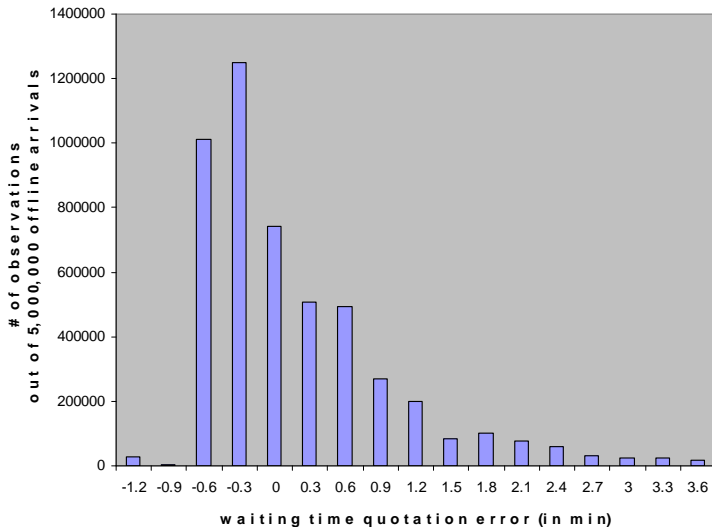


Figure 4: Histogram of the difference between the waiting time quotation and the actual waiting time in the case that $\alpha = 0.5$ in a simulation having Poisson arrivals with rate 100 per hour, deterministic service with mean 0.01, and parameters $\gamma = 0.01$, and $w_I = w_O = 1$

Recall that the system evolution equations in (3)-(9) make the simplifying assumption that the customers in the offline queue that 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 waiting 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 in order to ensure their presence when desired varies according to the system parameters and the average offline waiting time. Suppose we ask a customer that chooses to join the offline queue at time t to return to the service facility at time $\mathcal{W}_O(t) - \epsilon$. Table 5 shows what the value of ϵ must be in order to ensure 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. 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, and so our model is representative of the overall system behavior.

α	ϵ in min to achieve			Simulated avg. offline wait (in min)
	95%	98%	99%	
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

Table 5: The value of ϵ such that $\mathcal{W}_O(t) - \epsilon < W_O(t)$ for 95%, 98%, and 99% of the customers joining the offline queue for a simulation having Poisson arrivals with rate 100 per hour, deterministic service with mean 0.01 hours ($\mu = 100$), abandonment rate $\gamma = 0.01$, and $w_I = w_O = 1$.

5. The Amusement Park Ride Setting

An amusement park ride departs at deterministically spaced intervals, and can carry only a certain number of customers. Hence it is desirable to extend our analysis to include situations in which customers are served in batches at set time intervals. The parameter regime we have considered, in which the arrival and service rates are large, is applicable to most popular

amusement park rides, because generally hundreds or thousands of customers arrive to the ride, and board the ride, each hour.

The required modification to the model is the service process definition. In a slight abuse of notation, we use S_I^λ and S_O^λ to denote the cumulative number of customers that have boarded the ride from the inline and offline queues, even though the service processes are no longer renewal. The reader is to understand that in this Section, S_I^λ and S_O^λ refer to the processes defined below. Let

$$l^\lambda \equiv \left(\frac{1}{\lambda}\right)^{2/3},$$

and assume service occurs only at discrete time points $l^\lambda, 2l^\lambda, 3l^\lambda, \dots$, which represent ride departure times. At each discrete time il^λ , the number of customers that can enter into service (board the ride) is $n^\lambda \equiv \lambda^{1/3} - \lambda^{-1/6}\theta$. Then, the service rate is $\mu(\lambda) = n^\lambda/l^\lambda = \lambda - \sqrt{\lambda}\theta$ customers per hour.

The service process is defined recursively as follows. At time 0, no customers have boarded the ride, so that

$$S_I^\lambda(0) \equiv S_O^\lambda(0) \equiv 0.$$

Suppose that at discrete time il^λ , there are Q_I customers in the inline queue and Q_O customers in the offline queue. Then, the number of customers served from each queue that board the ride is

$$B_I^\lambda(il^\lambda) \equiv \begin{cases} \lfloor \alpha n^\lambda \rfloor + \min\left(\lceil [(1-\alpha)n^\lambda] - Q_O \rceil^+, Q_I - \lfloor \alpha n^\lambda \rfloor\right) & Q_I \geq \lfloor \alpha n^\lambda \rfloor \\ Q_I & Q_I < \lfloor \alpha n^\lambda \rfloor \end{cases}$$

and

$$B_O^\lambda(il^\lambda) \equiv \begin{cases} \lceil (1-\alpha)n^\lambda \rceil + \min\left(\lfloor \lfloor \alpha n^\lambda \rfloor - Q_I \rfloor^+, Q_O - \lceil (1-\alpha)n^\lambda \rceil\right) & Q_O \geq \lceil (1-\alpha)n^\lambda \rceil \\ Q_O & Q_O < \lceil (1-\alpha)n^\lambda \rceil \end{cases}.$$

Hence

$$\begin{aligned} S_I^\lambda(il^\lambda) &\equiv S_I^\lambda((i-1)l^\lambda) + B_I^\lambda(il^\lambda) \\ S_O^\lambda(il^\lambda) &\equiv S_O^\lambda((i-1)l^\lambda) + B_O^\lambda(il^\lambda). \end{aligned}$$

No customers board the ride in between the discrete time points $l^\lambda, 2l^\lambda, 3l^\lambda, \dots$, and so for any $t > 0$,

$$S_I^\lambda(t) = S_I^\lambda\left(\left\lfloor \frac{t}{l^\lambda} \right\rfloor l^\lambda\right) \text{ and } S_O^\lambda(t) = S_O^\lambda\left(\left\lfloor \frac{t}{l^\lambda} \right\rfloor l^\lambda\right).$$

The evolution equations for the queue-length process are very similar to (3) and (4) in Section 2

$$Q_I^\lambda(t) \equiv \sum_{i=1}^{A^\lambda(t)} \mathbf{1}\{w_I \mathcal{W}^\lambda(t_i^-) \leq w_O \mathcal{W}^\lambda(t_i^-)\} - S_I^\lambda(t) \quad (15)$$

$$Q_O^\lambda(t) \equiv \sum_{i=1}^{A^\lambda(t)} \mathbf{1}\{w_I \mathcal{W}^\lambda(t_i^-) > w_O \mathcal{W}^\lambda(t_i^-)\} - S_O^\lambda(t) - N \left(\int_0^t \gamma Q_O^\lambda(s) ds \right). \quad (16)$$

The difference is that now, because the service process counts the cumulative number of customers that have boarded the ride (and not the number of customers that have departed after riding), the processes Q_I^λ and Q_O^λ track only the customers waiting to ride (and do not include the customers riding or in service). Hence the wait time estimates

$$\mathcal{W}_I^\lambda(t) \equiv \frac{Q_I^\lambda(t)}{\mu(\lambda)\alpha} \text{ and } \mathcal{W}_O^\lambda(t) \equiv \frac{O^\lambda(t)}{\mu(\lambda)(1-\alpha)}$$

do not include any customers currently on the ride. This is reasonable because at time t the time remaining until the next ride departs is $(\lceil t/l^\lambda \rceil)l^\lambda - t$, which becomes negligible as λ increases. Finally, note that the bound on O^λ in (10) is now

$$Q_O^\lambda(t) \leq O^\lambda(t) \leq Q_O^\lambda(t) + N \left(\int_0^t \gamma Q_O^\lambda(s) ds \right) - N \left(\int_0^{\lceil t - \sup_{0 \leq s \leq t} W_O^\lambda(s) \rceil^+} \gamma Q_O^\lambda(s) ds \right), \quad (17)$$

where $W_O^\lambda(t)$ represents the actual time a customer arriving to the offline queue at time t must wait to board the ride.

We expect that the discrete review system behaves similarly to the continuous time system. The following proposition shows that Theorems 1 and 2, and hence also Corollary 1, remain valid for the discrete review system.

Proposition 1 *Theorems 1 and 2 remain valid for the model defined through (15)-(16). The process \tilde{Q} appearing in Theorems 1 and 2 again solves the stochastic equation (11) but the infinitesimal variance of the Brownian motion \tilde{X} is σ_A^2 .*

We end this Section by showing how to apply Proposition 1 to one popular roller coaster ride at Six Flags Magic Mountain, Tatsu. Suppose the arrival rate λ has been estimated. To use the approximation, we must determine the parameter θ . Tatsu has capacity for approximately 1600 people to ride every hour, and so

$$\lambda - \sqrt{\lambda}\theta = 1600, \text{ or } \theta = \frac{\lambda - 1600}{\sqrt{\lambda}}.$$

The approximation is not very sensitive to l^λ , because Proposition 1 remains valid for any review period of size λ^{-f} for $1/2 < f < 1$. The key is that the time between ride departures is roughly on the order of seconds if the number of people that can ride every hour is around one or two thousand (which is true for most roller coasters).

6. Conclusions

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 that 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 inline 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 based on Little's law despite the abandonments that may occur from the offline queue. Our results continue to hold in a setting that models an amusement park ride, in which customers are served in batches at discrete time points. We demonstrate the accuracy of our approximations via simulation.

Acknowledgments

We would like to thank Sriram Dasu for the original idea in this paper. We would also like to thank Mor Armony and Bob Foley for helpful discussions related to this paper. We would finally like to thank the associate editor and anonymous referees for their careful reading and useful suggestions that have very much improved this paper.

References

- Adan, I. J., Wessels, J., Zijm, W. H. M., 1991. Analysis of the asymmetric shortest queue problem. *Queueing Systems* **8**, 1–58.
- Ahmadi, R. H., 1997. Managing capacity and flow at theme parks. *Operations Research* **45**, 1–13.
- Armony, M., Maglaras, C., 2004a. Contact centers with a call-back option and real-time delay information. *Operations Research* **52**, 527–545.

- Armony, M., Maglaras, C., 2004b. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research* **52**, 271–292.
- Billingsley, P., 1999. *Convergence of Probability Measures*. John Wiley & Sons, Inc., New York, second Edition.
- Bitran, G. R., Ferrer, J. C., Oliveira, P. R., 2007. Managing customers experiences: Perspectives on the temporal aspects of service encounters. Forthcoming in the *Manufacturing & Service Operations Management*.
- Browne, S., Whitt, W., 1995. Piecewise-linear diffusion processes. In: Dshalalow, J. (Ed.), *Advances in Queueing: Theory, Methods, and Open Problems*. CRC Press, Boca Raton, Florida, pp. 463–480.
- Dickson, D., Ford, R. C., Laval, B., 2005. Managing real and virtual wait in hospitality and service organizations. *Cornell Hotel and Restaurant Administration Quarterly* **46**, 52–68.
- Flatto, L., McLean, H. P., 1977. Two queue in parallel. *Communications in Pure and Applied Mathematics* **30**, 255–263.
- Harrison, J. M., 1985. *Brownian Motion and Stochastic Flow Systems*. Krieger, Malabar, Florida.
- Iglehart, D. L., Whitt, W., 1970. Multiple channels queues in heavy traffic I. *Adv. in Applied Probability* **2**, 150–177.
- Katz, K., Larson, B., Larson, R., 1991. Prescription for the waiting in line bluse: Entertain, enlighten and engage. *Sloan Management Review* Winter, 44–53.
- Kruk, L., Lehoczky, J., Ramanan, K., Shreve, S., 2007. An explicit formula for the skorokhod map on $[0, a]$. *Annals of Probability* **35** (5), 1740–1768.
- Maister, D., 1985. The psychology of waiting in lines. In: Czepiel, J. A., Solomon, M., Surprenant, C. S. (Eds.), *The service encounter*. Lexington Books, Lexington, MA, pp. 113–123.
- Mandelbaum, A., Shimkin, N., 2000. A model for rational abandonments from invisible queues. *Queueing Systems* **36**, 141–173.
- Mielke, R., Zahralddin, A., Padam, D., Mastaglio, T., 1998. Simulation applied to theme park management. In: Medeiros, D. J., Watson, E. F., Carson, J. S., Manivannan, M. S. (Eds.), *Proceedings of the 1998 Winter Simulation Conference*. pp. 1199–1203.

- Munichor, N., Rafaeli, A., 2007. Number of apologies? Customer reactions to telephone waiting time fillers. *Journal of Applied Probability* **92**, 511–518.
- Oi, W. I., 1971. A disneyland dilemma: Two-part tariffs for a mickey mouse monopoly. *The Quarterly Journal of Economics* **85** (1), 77–96.
- Parlakturk, A., Kumar, S., 2004. Self-interested routing in queueing networks. *Management Science* **50** (7), 949–967.
- Plambeck, E., Kumar, S., Harrison, J. M., 2001. Asymptotic optimality of a single server queueing system with constraints on throughput times. *Queueing Systems* **39**, 23–54.
- Puhalskii, A., 1994. On the invariance principle for the first passage time. *Mathematics of Operations Research* **19**, 946–954.
- Reed, J., Ward, A. R., 2007. Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. Forthcoming in *Mathematics of Operations Research*.
- Reiman, M. I., 1984. Some diffusion approximations with state space collapse. In: Bacceli, F., Fayolle, G. (Eds.), *Modelling and Performance Evaluation Methodology*. Springer-Verlag, pp. 209–240.
- Taylor, S., 1994. Waiting for service: the relationship between delays and evaluations of service. *Journal of Marketing* **58**, 56–69.
- Tatsu Ride Statistics, 2007. <http://en.wikipedia.org/wiki/Tatsu>.
- Ward, A. R., Kumar, S., 2008. Asymptotically optimal control of a queue with impatient customers. *Mathematics of Operations Research* **33** (1), 167–202.
- Whitt, W., 1999. Improving service by informing customers about anticipated delays. *Management Science* **45**, 192–207.
- Whitt, W., 2002. *Stochastic-Process Limits*. Springer, New York.
- Whitt, W., 2006. Fluid models for multiserver queues with abandonments. *Operations Research* **54** (1), 37–54.

Managing Service Systems with an Offline Waiting Option and Customer Abandonment: Technical Appendix

Vasiliki Kostami • Amy R. Ward

July 23, 2008

In this technical appendix we provide the proofs for Theorem 1, Corollary 1, Theorem 2, and Proposition 1 stated in the manuscript titled: “Managing Service Systems with an Offline Waiting Option and Customer Abandonment”. This requires several Lemmas, which we state upfront, but whose proofs we defer to the end of this technical appendix.

Recall that we are considering a system in which the arrival rate λ becomes large, and the service rate is $\lambda - \sqrt{\lambda}\theta$ for some $\theta \in \mathfrak{R}$. We superscript any process or quantity associated with the system having arrival rate λ and service rate $\mu(\lambda) = \lambda - \sqrt{\lambda}\theta$ by λ . The numbering of all the equations in this technical appendix begins after the number of the last equation, (17), in the main paper body.

Note that we require the following technicalities. All random variables are defined on a common probability space (Ω, \mathcal{F}, P) . For each positive integer d , let D be the space of right continuous functions with left limits (RCLL) in \mathfrak{R}^d having time domain $[0, \infty)$. We endow D with the usual Skorokhod J_1 topology, and let M^d denote the Borel σ -algebra associated with the J_1 topology. All stochastic processes are measurable functions from (Ω, \mathcal{F}, P) into (D, M^d) for some appropriate dimension d . Suppose $\{\xi^n\}_{n=1}^\infty$ is a sequence of stochastic processes. The notation $\xi^n \Rightarrow \xi$ means that the probability measures induced by the ξ^n 's on (D, M^d) converge weakly to the probability measure on (D, M^d) induced by the stochastic process ξ . Note that we suppress d from the notation unless necessary. We often reference the functional strong law of large numbers, the functional central limit theorem, and the continuous mapping theorem. A convenient reference for these theorems is Billingsley (1999) or Whitt (2002). We use the notation e to denote the identity process $e(t) = t$ for all $t \geq 0$. We let “a.s.” denote “almost surely” and “u.o.c.” denote “uniformly on compact sets”.

It is useful to work with the system processes under law of large numbers (fluid) and central limit theorem (diffusion) scaling. Define the fluid scaled quantities

$$\begin{aligned}
\bar{A}^\lambda(t) &\equiv \frac{1}{\lambda}A^\lambda(t) - t \\
\bar{S}_I^\lambda(t) &\equiv \frac{1}{\lambda}S_I^\lambda(t) - (1 - \frac{\theta}{\sqrt{\lambda}})t \\
\bar{S}_O^\lambda(t) &\equiv \frac{1}{\lambda}S_O^\lambda(t) - (1 - \frac{\theta}{\sqrt{\lambda}})t \\
\bar{N}^\lambda(t) &\equiv \frac{1}{\lambda}N(\lambda t) - t \\
\bar{Q}_I^\lambda(t) &\equiv \frac{1}{\lambda}Q_I^\lambda(t) \\
\bar{Q}_O^\lambda(t) &\equiv \frac{1}{\lambda}Q_O^\lambda(t) \\
\bar{Q}^\lambda(t) &\equiv \frac{1}{\lambda}Q^\lambda(t) \\
\bar{\tau}^\lambda(t) &\equiv \frac{1}{\lambda} \int_0^t \gamma[Q_O^\lambda(s) - 1]^+ ds,
\end{aligned}$$

and the diffusion scaled quantities

$$\begin{aligned}
\tilde{A}^\lambda(t) &\equiv \sqrt{\lambda}(\frac{1}{\lambda}A^\lambda(t) - t) \\
\tilde{S}_I^\lambda(t) &\equiv \sqrt{\lambda}(\frac{1}{\lambda}S_I^\lambda(t) - (1 - \frac{\theta}{\sqrt{\lambda}})t) \\
\tilde{S}_O^\lambda(t) &\equiv \sqrt{\lambda}(\frac{1}{\lambda}S_O^\lambda(t) - (1 - \frac{\theta}{\sqrt{\lambda}})t) \\
\tilde{N}^\lambda(t) &\equiv \sqrt{\lambda}(\frac{1}{\lambda}N(\lambda t) - t) \\
\tilde{Q}_I^\lambda(t) &\equiv \frac{1}{\sqrt{\lambda}}Q_I^\lambda(t) \\
\tilde{Q}_O^\lambda(t) &\equiv \frac{1}{\sqrt{\lambda}}Q_O^\lambda(t) \\
\tilde{Q}^\lambda(t) &\equiv \frac{1}{\sqrt{\lambda}}Q^\lambda(t) \\
\tilde{I}^\lambda(t) &\equiv \sqrt{\lambda}(1 - \frac{\theta}{\sqrt{\lambda}})I^\lambda(t) \\
\tilde{\mathcal{W}}_I^\lambda(t) &\equiv \sqrt{\lambda}\mathcal{W}_I^\lambda(t) \\
\tilde{\mathcal{W}}_O^\lambda(t) &\equiv \sqrt{\lambda}\mathcal{W}_O^\lambda(t) \\
\tilde{W}_I^\lambda &\equiv \sqrt{\lambda}W_I^\lambda \\
\tilde{W}_O^\lambda &\equiv \sqrt{\lambda}W_O^\lambda
\end{aligned}$$

It is additionally useful to introduce the processes P_I^λ and P_O^λ , which represent the workload processes in the inline and offline queues respectively. We use the term “workload” to indicate the total processing time of all the customers in the queue that will eventually receive service when all the effort of the server is given exclusively to their queue ($\alpha = 1$ or 0). Note that the workload process is defined conditionally on future abandonments, because the wait time of a customer is not affected by the customers in front of him that abandon. In particular, the actual waiting times a customer arriving to the inline or offline queue at time t would experience, $W_I^\lambda(t)$ and $W_O^\lambda(t)$, are exactly $P_I^\lambda(t)/\alpha$ and $P_O^\lambda(t)/(1-\alpha)$ when the server works continuously at rate α on the inline queue and at rate $(1-\alpha)$ on the offline queue. Also define the diffusion-scaled workload processes $\tilde{P}_I^\lambda = \sqrt{\lambda}P_I^\lambda$ and $\tilde{P}_O^\lambda = \sqrt{\lambda}P_O^\lambda$.

Next, we will use the following four Lemmas, whose proofs we defer to the end of the appendix.

Lemma 1 *As $\lambda \rightarrow \infty$,*

$$(\bar{Q}^\lambda, P_I^\lambda, P_O^\lambda, \bar{\tau}^\lambda, T_I^\lambda + T_O^\lambda, I^\lambda) \rightarrow (0, 0, 0, 0, e, 0), \text{ a.s., u.o.c..}$$

It is useful to observe that as a consequence of Lemma 1

$$\tilde{N}^\lambda \circ \bar{\tau}^\lambda \Rightarrow 0, \tag{18}$$

as $\lambda \rightarrow \infty$. This weak convergence follows because the functional central limit theorem establishes that \tilde{N}^λ weakly converges to a Brownian motion as $\lambda \rightarrow \infty$. Since the initial position of the Brownian motion is 0, and $\bar{\tau}^\lambda$ is a non-decreasing process, the random time change theorem establishes (18).

Lemma 2 *The sequence $\{\tilde{Q}^\lambda, \tilde{I}^\lambda\}$ is tight in D .*

Lemma 3 *Consider a system having arrival rate λ and service rate $\mu(\lambda) \equiv \lambda - \sqrt{\lambda}\theta$ for some $\theta \in \mathfrak{R}$. For \tilde{Q} defined by (11) in which \tilde{X} is a Brownian motion with infinitesimal drift θ and infinitesimal variance σ^2 ,*

$$\sqrt{\lambda}W_I^\lambda \Rightarrow \frac{w_O}{(1-\alpha)w_I + \alpha w_O} \tilde{Q} \text{ and } \sqrt{\lambda}W_O^\lambda \Rightarrow \frac{w_I}{(1-\alpha)w_I + \alpha w_O} \tilde{Q}, \text{ as } \lambda \rightarrow \infty. \tag{19}$$

Furthermore, for any $T > 0$, as $\lambda \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \left| \tilde{W}_I^\lambda(t) - \frac{\tilde{P}_I^\lambda(t)}{\alpha} \right| \rightarrow 0 \text{ and } \sup_{0 \leq t \leq T} \left| \tilde{W}_O^\lambda(t) - \frac{\tilde{P}_O^\lambda(t)}{1-\alpha} \right| \rightarrow 0, \text{ in probability.} \tag{20}$$

Our final Lemma states that Lemmas 1-3 remain valid in the modified model in Section 5, in which customers are served in batches at set time intervals. In this setting, as is true for the processes Q_I^λ and Q_O^λ , the workload processes P_I^λ and P_O^λ , and the actual waiting time processes W_I^λ and W_O^λ , refer to the customers waiting to board the ride (and do not include the customers currently riding). Furthermore,

$$\bar{\tau}^\lambda(t) \equiv \frac{1}{\lambda} \int_0^t \gamma Q_O^\lambda(s) ds.$$

Lemma 4 *Lemmas 1-3 also hold when the system evolution equations are specified through (15)-(16). Note that in this setting the processes T_I^λ , T_O^λ , and I^λ no longer appear in the system evolution equations, and so Lemma 1 is modified to state that as $\lambda \rightarrow \infty$,*

$$(\bar{Q}^\lambda, P_I^\lambda, P_O^\lambda, \bar{\tau}^\lambda) \rightarrow (0, 0, 0, 0), \text{ a.s., u.o.c..} \quad (21)$$

Finally, it is useful to write the stochastic equation for the diffusion \tilde{Q} in (11) in terms of the one-sided linearly generalized regulator mapping, whose definition we provide below.

Definition 1 *(The one-sided linearly generalized regulator mapping)*

Given κ a non-negative constant and $x \in D([0, \infty), \mathfrak{R})$ having $x(0) \geq 0$, the one-sided linearly generalized regulator mapping

$$(\phi^\kappa, \psi^\kappa) : D([0, \infty), \mathfrak{R}) \mapsto D([0, \infty), [0, \infty) \times [0, \infty))$$

is defined by

$$(\phi^\kappa, \psi^\kappa)(x) \equiv (z, l)$$

where

$$(C1) \ z(t) = x(t) - \kappa \int_0^t z(s) ds + l(t) \in [0, \infty) \text{ for all } t \geq 0;$$

$$(C2) \ l \text{ is nondecreasing, } l(0) = 0, \text{ and } \int_0^\infty z(t) dl(t) = 0.$$

Specifically, for

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

it follows that

$$(\tilde{Q}, \tilde{I}) = (\phi^\kappa, \psi^\kappa) (\tilde{X}). \quad (22)$$

Proposition 4.1 part (i) in Reed and Ward (2007) establishes the existence and uniqueness of the regulator mapping in Definition 1¹, and so the representation (22) is equivalent to the representation for \tilde{Q} in (11).

Proof of Theorem 1

Proof of (i)

The structure of our proof follows the proof of Theorem 1 in Section 5 in Reiman (1984), which establishes state-space collapse for a join the shorter queue system in heavy traffic with no abandonments. However, more delicate argument is required to handle the customer abandonments.

We need to show that for any $\epsilon > 0$,

$$P \left(\sup_{0 \leq t \leq T} \left| \frac{w_I}{\alpha} \tilde{Q}_I^\lambda(t) - \frac{w_O}{1-\alpha} \tilde{Q}_O^\lambda(t) \right| > \epsilon \right) \rightarrow 0 \text{ as } \lambda \rightarrow \infty. \quad (23)$$

Fix $\epsilon > 0$ and let

$$\begin{aligned} \xi_\lambda &\equiv \inf \left\{ t \geq 0 : \left| \frac{w_I}{\alpha} \tilde{Q}_I^\lambda(t) - \frac{w_O}{1-\alpha} \tilde{Q}_O^\lambda(t) \right| > \epsilon \right\} \\ \xi_\lambda^* &\equiv \sup \left\{ t \leq \xi_\lambda : \left| \frac{w_I}{\alpha} \tilde{Q}_I^\lambda(t) - \frac{w_O}{1-\alpha} \tilde{Q}_O^\lambda(t) \right| \leq \frac{\epsilon}{2} \right\}. \end{aligned}$$

It will also be useful to define the processes

$$\begin{aligned} \tilde{U}_1^\lambda(t, s, u, v) &\equiv -\frac{w_I}{\alpha} \left\{ \tilde{S}_I^\lambda(u + \alpha(t-s)) - \tilde{S}_I^\lambda(u) \right\} \\ &\quad + \frac{w_O}{1-\alpha} \left\{ \tilde{S}_O^\lambda(v + (1-\alpha)(t-s)) - \tilde{S}_O^\lambda(v) \right\} \\ &\quad - \frac{w_O}{1-\alpha} \left\{ \tilde{A}^\lambda(t) - \tilde{A}^\lambda(s) \right\} \\ &\quad + \left\{ \frac{\theta}{\sqrt{\lambda}}(w_I - w_O) - \frac{\alpha}{1-\alpha} w_O - w_I \right\} \sqrt{\lambda}(t-s) \\ \tilde{U}_2^\lambda(t, s, u, v) &\equiv -\frac{w_O}{1-\alpha} \left\{ \tilde{S}_O^\lambda(v + (1-\alpha)(t-s)) - \tilde{S}_O^\lambda(v) \right\} \\ &\quad + \frac{w_I}{\alpha} \left\{ \tilde{S}_I^\lambda(u + \alpha(t-s)) - \tilde{S}_I^\lambda(u) \right\} \\ &\quad - \frac{w_I}{\alpha} \left\{ \tilde{A}^\lambda(t) - \tilde{A}^\lambda(s) \right\} \\ &\quad + \left\{ \frac{\theta}{\sqrt{\lambda}}(w_O - w_I) - \frac{(1-\alpha)}{\alpha} w_I - w_O \right\} \sqrt{\lambda}(t-s). \end{aligned}$$

¹Actually, the regulator mapping in Definition 1 is a specific instance of the more general regulator mapping in Reed and Ward (2007).

An upper bound for the left-hand-side of (23)

First assume $w_I \tilde{Q}_I^\lambda(\xi_\lambda^*)/\alpha > w_O \tilde{Q}_O^\lambda(\xi_\lambda^*)/(1-\alpha)$. Then, for $\xi_\lambda^* \leq t \leq \xi_\lambda$, all customers join the offline service queue, and so

$$\begin{aligned}
& \left| \frac{w_I}{\alpha} \tilde{Q}_I^\lambda(t) - \frac{w_O}{1-\alpha} \tilde{Q}_O^\lambda(t) \right| \\
&= \frac{w_I}{\alpha} \tilde{Q}_I^\lambda(\xi_\lambda^*-) - \frac{w_O}{1-\alpha} \tilde{Q}_O^\lambda(\xi_\lambda^*-) - \frac{w_I}{\alpha} \frac{1}{\sqrt{\lambda}} \{S_I^\lambda(T_I^\lambda(t)) - S_I^\lambda(T_I^\lambda(\xi_\lambda^*-))\} \\
&\quad + \frac{w_O}{1-\alpha} \frac{1}{\sqrt{\lambda}} \{S_O^\lambda(T_O^\lambda(t)) - S_O^\lambda(T_O^\lambda(\xi_\lambda^*-))\} - \frac{w_O}{1-\alpha} \left\{ \tilde{A}^\lambda(t) - \tilde{A}^\lambda(\xi_\lambda^*-) + \sqrt{\lambda}(t - \xi_\lambda^*) \right\} \\
&\quad + \frac{w_O}{1-\alpha} \frac{1}{\sqrt{\lambda}} \left[N \left(\int_0^t \gamma [Q_O^\lambda(s) - 1]^+ ds \right) - N \left(\int_0^{\xi_\lambda^*-} \gamma [Q_O^\lambda(s) - 1]^+ ds \right) \right]. \quad (24)
\end{aligned}$$

The inline queue does not become empty during $[\xi_\lambda^*, \xi_\lambda]$, so that

$$T_I^\lambda(t) - T_I^\lambda(\xi_\lambda^*-) \geq \alpha(t - \xi_\lambda^*).$$

The offline queue may become empty during $[\xi_\lambda^*, \xi_\lambda]$, so that

$$T_O^\lambda(t) - T_O^\lambda(\xi_\lambda^*-) \leq (1-\alpha)(t - \xi_\lambda^*).$$

Since S_I^λ and S_O^λ are non-decreasing processes,

$$\begin{aligned}
& S_I^\lambda(T_I^\lambda(t)) - S_I^\lambda(T_I^\lambda(\xi_\lambda^*-)) \\
&\geq S_I^\lambda(T_I^\lambda(\xi_\lambda^*-) + \alpha(t - \xi_\lambda^*)) - S_I^\lambda(T_I^\lambda(\xi_\lambda^*-)) \\
&= \sqrt{\lambda} \left[\tilde{S}_I^\lambda(T_I^\lambda(\xi_\lambda^*-) + \alpha(t - \xi_\lambda^*)) - \tilde{S}_I^\lambda(T_I^\lambda(\xi_\lambda^*-)) + \alpha \left(1 - \frac{\theta}{\sqrt{\lambda}} \right) \sqrt{\lambda}(t - \xi_\lambda^*) \right],
\end{aligned}$$

and

$$\begin{aligned}
& S_O^\lambda(T_O^\lambda(t)) - S_O^\lambda(T_O^\lambda(\xi_\lambda^*-)) \\
&\leq S_O^\lambda(T_O^\lambda(\xi_\lambda^*-) + (1-\alpha)(t - \xi_\lambda^*)) - S_O^\lambda(T_O^\lambda(\xi_\lambda^*-)) \\
&= \sqrt{\lambda} \left[\tilde{S}_O^\lambda(T_O^\lambda(\xi_\lambda^*-) + (1-\alpha)(t - \xi_\lambda^*)) - \tilde{S}_O^\lambda(T_O^\lambda(\xi_\lambda^*-)) + (1-\alpha) \left(1 - \frac{\theta}{\sqrt{\lambda}} \right) \sqrt{\lambda}(t - \xi_\lambda^*) \right].
\end{aligned}$$

The definition of ξ_λ^* and substitution of the above upper bounds into (24) establish

$$\begin{aligned} & \left| \frac{w_I}{\alpha} \tilde{Q}_I^\lambda(t) - \frac{w_O}{1-\alpha} \tilde{Q}_O^\lambda(t) \right| \\ & \leq \frac{\epsilon}{2} + \tilde{U}_1^\lambda(t, \xi_\lambda^*, T_I^\lambda(\xi_\lambda^*), T_O^\lambda(\xi_\lambda^*)) \\ & + \frac{w_O}{1-\alpha} \frac{1}{\sqrt{\lambda}} \left[N \left(\int_0^t \gamma [Q_O^\lambda(s) - 1]^+ ds \right) - N \left(\int_0^{\xi_\lambda^{*-}} \gamma [Q_O^\lambda(s) - 1]^+ ds \right) \right]. \end{aligned}$$

When $w_I \tilde{Q}_I^\lambda(\xi_\lambda^*) / \alpha \leq w_O \tilde{Q}_O^\lambda(\xi_\lambda^*) / (1-\alpha)$, a similar argument shows

$$\begin{aligned} & \left| \frac{w_O}{1-\alpha} \tilde{Q}_O^\lambda(t) - \frac{w_I}{\alpha} \tilde{Q}_I^\lambda(t) \right| \\ & \leq \frac{\epsilon}{2} + \tilde{U}_2^\lambda(t, \xi_\lambda^*, T_I^\lambda(\xi_\lambda^*), T_O^\lambda(\xi_\lambda^*)) \\ & - \frac{w_O}{1-\alpha} \frac{1}{\sqrt{\lambda}} \left[N \left(\int_0^t \gamma [Q_O^\lambda(s) - 1]^+ ds \right) - N \left(\int_0^{\xi_\lambda^{*-}} \gamma [Q_O^\lambda(s) - 1]^+ ds \right) \right]. \end{aligned}$$

Also noting the process N is non-negative, we conclude

$$\begin{aligned} & \left| \frac{w_I}{\alpha} \tilde{Q}_I^\lambda(t) - \frac{w_O}{1-\alpha} \tilde{Q}_O^\lambda(t) \right| \\ & \leq \frac{\epsilon}{2} + \max \left\{ \tilde{U}_1^\lambda(t, \xi_\lambda^*, T_I^\lambda(\xi_\lambda^*), T_O^\lambda(\xi_\lambda^*)), \tilde{U}_2^\lambda(t, \xi_\lambda^*, T_I^\lambda(\xi_\lambda^*), T_O^\lambda(\xi_\lambda^*)) \right\} \\ & + \frac{w_O}{1-\alpha} \frac{1}{\sqrt{\lambda}} \left(N \left(\int_0^t \gamma [Q_O^\lambda(w) - 1]^+ dw \right) - N \left(\int_0^{\xi_\lambda^{*-}} \gamma [Q_O^\lambda(w) - 1]^+ dw \right) \right). \end{aligned}$$

Therefore, the left-hand side of (23) can be bounded as follows

$$\begin{aligned} & P \left(\sup_{0 \leq t \leq T} \left| \frac{w_I}{\alpha} \tilde{Q}_I^\lambda(t) - \frac{w_O}{1-\alpha} \tilde{Q}_O^\lambda(t) \right| > \epsilon \right) \\ & \leq P \left(\sup_{0 \leq s \leq t \leq T} \sup_{0 \leq u, v \leq s} \max \left\{ \tilde{U}_1^\lambda(t, s, u, v), \tilde{U}_2^\lambda(t, s, u, v) \right\} \right. \\ & \quad \left. + \frac{w_O}{1-\alpha} \frac{1}{\sqrt{\lambda}} \left(N \left(\int_0^t \gamma [Q_O^\lambda(w) - 1]^+ dw \right) - N \left(\int_0^s \gamma [Q_O^\lambda(w) - 1]^+ dw \right) \right) > \frac{\epsilon}{2} \right). \end{aligned} \tag{25}$$

Convergence of the right-hand-side of (25) to zero

Let η be arbitrarily small. We will show that the right-hand side of (25) is less than η for large enough λ .

Define

$$\begin{aligned}
\tilde{B}_I^\lambda(t) &\equiv \sup_{0 \leq s \leq t, 0 \leq u, v \leq s} -\frac{w_I}{\alpha} \left\{ \tilde{S}_I^\lambda(u + \alpha(t-s)) - \tilde{S}_I^\lambda(u) \right\} \\
&\quad + \frac{w_O}{1-\alpha} \left\{ \tilde{S}_O^\lambda(v + (1-\alpha)(t-s)) - \tilde{S}_O^\lambda(v) \right\} - \frac{w_O}{1-\alpha} \left\{ \tilde{A}^\lambda(t) - \tilde{A}^\lambda(s) \right\} \\
\tilde{B}_O^\lambda(t) &\equiv \sup_{0 \leq s \leq t, 0 \leq u, v \leq s} -\frac{w_O}{1-\alpha} \left\{ \tilde{S}_O^\lambda(v + (1-\alpha)(t-s)) - \tilde{S}_O^\lambda(v) \right\} \\
&\quad + \frac{w_I}{\alpha} \left\{ \tilde{S}_I^\lambda(u + \alpha(t-s)) - \tilde{S}_I^\lambda(u) \right\} - \frac{w_I}{\alpha} \left\{ \tilde{A}^\lambda(t) - \tilde{A}^\lambda(s) \right\}.
\end{aligned}$$

Also observe that

$$\begin{aligned}
&\max \left(\tilde{U}_1^\lambda(t, s, u, v), \tilde{U}_2^\lambda(t, s, u, v) \right) \\
&\leq \max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \sqrt{\lambda}(t-s) \max \left(\frac{\theta}{\sqrt{\lambda}}(w_I - w_O) - \frac{\alpha}{1-\alpha}w_O - w_I, \right. \\
&\quad \left. \frac{\theta}{\sqrt{\lambda}}(w_O - w_I) - \frac{1-\alpha}{\alpha}w_I - w_O \right)
\end{aligned}$$

(because for any constants d_1, d_2, d_3 , and d_4 , $\max(d_1 + d_2, d_3 + d_4) \leq \max(d_1, d_3) + \max(d_2, d_4)$). Furthermore, for $\underline{w} \equiv \min(w_I, w_O)$,

$$\max \left(\frac{\theta}{\sqrt{\lambda}}(w_I - w_O) - \frac{\alpha}{1-\alpha}w_O - w_I, \frac{\theta}{\sqrt{\lambda}}(w_O - w_I) - \frac{1-\alpha}{\alpha}w_I - w_O \right) \leq \frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w},$$

and so

$$\max \left(\tilde{U}_1^\lambda(t, s, u, v), \tilde{U}_2^\lambda(t, s, u, v) \right) \leq \max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \sqrt{\lambda}(t-s) \left(\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} \right).$$

Next, for any $t > 0$,

$$\frac{1}{\sqrt{\lambda}}N \left(\int_0^t \gamma [Q_O^\lambda(w) - 1]^+ dw \right) = \tilde{N}^\lambda(\bar{\tau}^\lambda(t)) + \gamma \int_0^t \frac{[Q_O^\lambda(w) - 1]^+}{\sqrt{\lambda}} dw,$$

and so

$$\begin{aligned}
&\frac{1}{\sqrt{\lambda}} \left(N \left(\int_0^t \gamma [Q_O^\lambda(w) - 1]^+ dw \right) - N \left(\int_0^s \gamma [Q_O^\lambda(w) - 1]^+ dw \right) \right) \\
&= \tilde{N}^\lambda(\bar{\tau}(t)) - \tilde{N}^\lambda(\bar{\tau}(s)) + \gamma \int_s^t \frac{[Q_O^\lambda(w) - 1]^+}{\sqrt{\lambda}} dw \\
&\leq \left| \tilde{N}^\lambda(\bar{\tau}^\lambda(t)) \right| + \left| \tilde{N}^\lambda(\bar{\tau}^\lambda(s)) \right| + \gamma(t-s) \sup_{0 \leq t \leq T} \tilde{Q}^\lambda(t).
\end{aligned}$$

We conclude that

$$\begin{aligned}
& P \left(\begin{aligned} & \max \left\{ \tilde{U}_1^\lambda(t, s, u, v), \tilde{U}_2^\lambda(t, s, u, v) \right\} \\ & \sup_{0 \leq s \leq t \leq T} \sup_{0 \leq u, v \leq s} + \frac{w_O}{1-\alpha} \frac{1}{\sqrt{\lambda}} \left(\begin{aligned} & N \left(\int_0^t \gamma [Q_O^\lambda(w) - 1]^+ dw \right) \\ & -N \left(\int_0^s \gamma [Q_O^\lambda(w) - 1]^+ dw \right) \end{aligned} \right) > \frac{\epsilon}{2} \end{aligned} \right) \quad (26) \\
& \leq P \left(\begin{aligned} & \sup_{0 \leq s \leq t \leq T} \max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \left| \tilde{N}^\lambda(\bar{\tau}^\lambda(t)) \right| + \left| \tilde{N}^\lambda(\bar{\tau}^\lambda(s)) \right| \\ & + \gamma(t-s) \sup_{0 \leq t \leq T} \tilde{Q}^\lambda(t) + \sqrt{\lambda}(t-s) \left(\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} \right) > \frac{\epsilon}{2} \end{aligned} \right).
\end{aligned}$$

From (18), $\tilde{N}^\lambda(\bar{\tau}^\lambda(T)) \Rightarrow 0$ as $\lambda \rightarrow \infty$. Hence for large enough λ

$$P \left(\sup_{0 \leq t \leq T} \tilde{N}^\lambda(\bar{\tau}^\lambda(t)) > \frac{\epsilon}{8} \right) < \frac{\eta}{4},$$

and so

$$\begin{aligned}
& P \left(\begin{aligned} & \sup_{0 \leq s \leq t \leq T} \max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \left| \tilde{N}^\lambda(\bar{\tau}^\lambda(t)) \right| + \left| \tilde{N}^\lambda(\bar{\tau}^\lambda(s)) \right| \\ & + \gamma(t-s) \sup_{0 \leq t \leq T} \tilde{Q}^\lambda(t) + \sqrt{\lambda}(t-s) \left(\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} \right) > \frac{\epsilon}{2} \end{aligned} \right) \\
& \leq P \left(\begin{aligned} & \sup_{0 \leq s \leq t \leq T} \left(\begin{aligned} & \max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \gamma(t-s) \sup_{0 \leq t \leq T} \tilde{Q}^\lambda(t) \\ & + \sqrt{\lambda}(t-s) \left(\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} \right) \end{aligned} \right) > \frac{\epsilon}{4} \end{aligned} \right) + \frac{\eta}{4}.
\end{aligned}$$

Furthermore, it follows from Lemma 2 that there exists a finite positive constant M such that

$$P \left(\sup_{0 \leq t \leq T} \tilde{Q}^\lambda(t) > M \right) < \frac{\eta}{4}$$

for all large enough λ . Therefore,

$$\begin{aligned}
& P \left(\begin{aligned} & \sup_{0 \leq s \leq t \leq T} \max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \gamma(t-s) \sup_{0 \leq t \leq T} \tilde{Q}^\lambda(t) + \sqrt{\lambda}(t-s) \left(\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} \right) > \frac{\epsilon}{4} \end{aligned} \right) \\
& \leq P \left(\begin{aligned} & \sup_{0 \leq s \leq t \leq T} \max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \gamma(t-s)M + \sqrt{\lambda}(t-s) \left(\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} \right) > \frac{\epsilon}{4} \end{aligned} \right) + \frac{\eta}{4}.
\end{aligned}$$

We conclude that

$$\begin{aligned}
& P \left(\begin{aligned} & \sup_{0 \leq s \leq t \leq T} \max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \left| \tilde{N}^\lambda(\bar{\tau}^\lambda(t)) \right| + \left| \tilde{N}^\lambda(\bar{\tau}^\lambda(s)) \right| \\ & + \gamma(t-s) \sup_{0 \leq t \leq T} \tilde{Q}^\lambda(t) + \sqrt{\lambda}(t-s) \left(\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} \right) > \frac{\epsilon}{2} \end{aligned} \right) \\
& \leq P \left(\begin{aligned} & \sup_{0 \leq s \leq t \leq T} \max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \gamma(t-s)M + \sqrt{\lambda}(t-s) \left(\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} \right) > \frac{\epsilon}{4} \end{aligned} \right) + \frac{\eta}{4}.
\end{aligned}$$

Hence, from (26), to complete the proof, it is sufficient to show that

$$P \left(\sup_{0 \leq s \leq t \leq T} \left(\max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \gamma(t-s)M + \sqrt{\lambda}(t-s) \left(\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} \right) \right) > \frac{\epsilon}{4} \right) < \frac{\eta}{2}. \quad (27)$$

This argument is similar to the paragraph following (10) in the proof of Theorem 3.2 in Reiman (1984), and we show the details in the next two paragraphs for the reader's convenience.

Details of the argument that (27) holds

For large enough λ , since $w_I > 0$ and $w_O > 0$,

$$\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} < 0.$$

For any $\nu \in (0, T)$, for large enough λ ,

$$\begin{aligned} & \sup_{0 \leq s \leq t \leq T} \max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \gamma(t-s)M + \sqrt{\lambda}(t-s) \left(\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} \right) \\ & \leq \max \left(\tilde{B}_I^\lambda(\nu), \tilde{B}_O^\lambda(\nu) \right) + \gamma\nu M + \max \left(\tilde{B}_I^\lambda(T), \tilde{B}_O^\lambda(T) \right) + \gamma TM + \sqrt{\lambda}\nu \left(\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} \right). \end{aligned}$$

Hence

$$\begin{aligned} & P \left(\sup_{0 \leq s \leq t \leq T} \max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \gamma(t-s)M + \sqrt{\lambda}(t-s) \left(\frac{\theta}{\sqrt{\lambda}}(w_I + w_O) - \underline{w} \right) > \frac{\epsilon}{4} \right) \\ & \leq P \left(\max \left(\tilde{B}_I^\lambda(\nu), \tilde{B}_O^\lambda(\nu) \right) + \gamma\nu M > \frac{\epsilon}{8} \right) \\ & \quad + P \left(\max \left(\tilde{B}_I^\lambda(T), \tilde{B}_O^\lambda(T) \right) + \gamma TM > \frac{\epsilon}{8} + \sqrt{\lambda}\nu \left(\underline{w} - \frac{\theta}{\sqrt{\lambda}}(w_I + w_O) \right) \right). \quad (28) \end{aligned}$$

The functional central limit theorem shows that \tilde{S}_O^λ , \tilde{S}_I^λ , and \tilde{A}^λ all weakly converge to mean 0 Brownian motions. Hence $\tilde{B}_I^\lambda(t) \Rightarrow 0$ and $\tilde{B}_O^\lambda(t) \Rightarrow 0$ as $t \rightarrow \infty$. This means that we can choose ν small enough so that

$$P \left(\max \left(\tilde{B}_I^\lambda(\nu), \tilde{B}_O^\lambda(\nu) \right) + \gamma\nu M > \frac{\epsilon}{8} \right) < \frac{\eta}{4}.$$

It is also true that $\tilde{B}_I^\lambda(T)$ and $\tilde{B}_O^\lambda(T)$ weakly converge to mean 0 normal random variables.

Hence we can choose λ large enough so that

$$P \left(\max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \gamma TM > \frac{\epsilon}{8} + \sqrt{\lambda} \nu \left(\underline{w} - \frac{\theta}{\sqrt{\lambda}} (w_I + w_O) \right) \right) < \frac{\eta}{4}.$$

Then, it follows from (28) that

$$P \left(\sup_{0 \leq s \leq t \leq T} \max \left(\tilde{B}_I^\lambda(t), \tilde{B}_O^\lambda(t) \right) + \gamma(t-s)M + \sqrt{\lambda}(t-s) \left(\frac{\theta}{\sqrt{\lambda}} (w_I + w_O) - \underline{w} \right) > \frac{\epsilon}{4} \right) < \frac{\eta}{2},$$

which shows that (27) is valid, and so completes the proof.

Proof of (ii)

We first represent \tilde{Q}^λ using the one-sided linearly generalized regulator mapping given in Definition 1. Let

$$\begin{aligned} \tilde{X}^\lambda(t) &\equiv \tilde{A}^\lambda(t) - \tilde{S}_I^\lambda(T_I^\lambda(t)) - \tilde{S}_O^\lambda(T_O^\lambda(t)) - \tilde{N}^\lambda(\bar{\tau}^\lambda(t)) + \theta t \\ \tilde{\epsilon}^\lambda(t) &\equiv \gamma \int_0^t \left(\frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} \tilde{Q}^\lambda(s) - \left[\tilde{Q}_O^\lambda(s) - \frac{1}{\sqrt{\lambda}} \right]^+ \right) ds. \end{aligned}$$

Then, for all $t \geq 0$,

$$\tilde{Q}^\lambda(t) = \tilde{X}^\lambda(t) + \tilde{\epsilon}^\lambda(t) - \gamma \int_0^t \frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} \tilde{Q}^\lambda(s) + \tilde{I}^\lambda(t) \geq 0.$$

Since also \tilde{I}^λ is non-decreasing, $\tilde{I}^\lambda(0) = 0$, and

$$\int_0^\infty \tilde{Q}^\lambda(t) d\tilde{I}^\lambda(t) = \int_0^\infty Q^\lambda(t) \mathbf{1}\{Q^\lambda(t) = 0\} dt = 0,$$

it follows that

$$\left(\tilde{Q}^\lambda, \tilde{I}^\lambda \right) \equiv (\phi^\kappa, \psi^\kappa) \left(\tilde{X}^\lambda + \tilde{\epsilon}^\lambda \right). \quad (29)$$

Suppose we can show

$$\tilde{X}^\lambda \Rightarrow \tilde{X}, \quad (30)$$

as $\lambda \rightarrow \infty$, where \tilde{X} is a Brownian motion with drift θ and variance $\sigma^2 = \sigma_A^2 + \sigma_S^2$ as in (11).

Suppose we can also show that

$$\tilde{\epsilon}^\lambda \Rightarrow 0, \quad (31)$$

as $\lambda \rightarrow \infty$. Proposition 4.1 part (iv) in Reed and Ward (2007) establishes that the mapping

$(\phi^\kappa, \psi^\kappa)$ is continuous. Therefore, by the continuous mapping theorem

$$(\phi^\kappa, \psi^\kappa) \left(\tilde{X}^\lambda + \tilde{\varepsilon}^\lambda \right) \Rightarrow (\phi^\kappa, \psi^\kappa) \left(\tilde{X} \right),$$

as $\lambda \rightarrow \infty$. The representation (\tilde{Q}, \tilde{I}) in terms of the one-sided linearly generalized regulator mapping shows $(\tilde{Q}, \tilde{I}) = (\phi^\kappa, \psi^\kappa)(\tilde{X})$, and so

$$\left(\tilde{Q}^\lambda, \tilde{I}^\lambda \right) \Rightarrow (\tilde{Q}, \tilde{I})$$

as $\lambda \rightarrow \infty$.

We now establish (30). The sequence $\{(T_O^\lambda, T_I^\lambda)\}$ is tight in D because $|T_I^\lambda(t) - T_I^\lambda(s)| \leq |t - s|$ and $|T_O^\lambda(t) - T_O^\lambda(s)| \leq |t - s|$. Consider any subsequence $\{\lambda_k\}$ on which

$$\left(T_O^{\lambda_k}, T_I^{\lambda_k} \right) \Rightarrow (T_O, T_I)$$

as $\lambda_k \rightarrow \infty$. By Lemma 1, the limit process satisfies

$$T_O + T_I = e.$$

Let B_1, B_2 , and B_3 be independent, standard Brownian motions. On the subsequence $\{\lambda_k\}$, by the functional central limit theorem and the continuous mapping theorem

$$\tilde{A}^{\lambda_k} - \tilde{S}_I^{\lambda_k} \circ T_I^{\lambda_k} - \tilde{S}_O^{\lambda_k} \circ T_O^{\lambda_k} + \theta e \Rightarrow \sigma_A B_1 - \sigma_S B_2 \circ T_I - \sigma_S B_3 \circ T_O + \theta e,$$

as $\lambda_k \rightarrow \infty$. By (18)

$$\tilde{N}^{\lambda_k} \circ \bar{\tau}^{\lambda_k} \Rightarrow 0,$$

as $\lambda_k \rightarrow \infty$. Therefore,

$$\tilde{X}^{\lambda_k} \Rightarrow \sigma_A B_1 - \sigma_S B_2 \circ T_I - \sigma_S B_3 \circ T_O + \theta e \stackrel{D}{=} \tilde{X},$$

as $\lambda_k \rightarrow \infty$, where the symbol $\stackrel{D}{=}$ denotes equality in distribution. Since the subsequence $\{\lambda_k\}$ was arbitrary, we conclude

$$\tilde{X}^\lambda \Rightarrow \tilde{X},$$

as $\lambda \rightarrow \infty$.

Finally, to establish (31) and complete the proof, let λ_k be a subsequence along which

$$\tilde{Q}^{\lambda_k} \Rightarrow \tilde{Q}$$

as $\lambda_k \rightarrow \infty$. Such a subsequence exists by Lemma 2. On this subsequence, by part (i) and the fact that $\tilde{Q}^{\lambda_k} = \tilde{Q}_I^{\lambda_k} + \tilde{Q}_O^{\lambda_k}$, for any $T > 0$,

$$\sup_{0 \leq t \leq T} \frac{1}{\sqrt{\lambda_k}} \left| \frac{w_I}{\alpha} \left(Q^{\lambda_k}(t) - Q_O^{\lambda_k}(t) \right) - \frac{w_O}{1-\alpha} Q_O^{\lambda_k}(t) \right| \rightarrow 0, \text{ in probability as } \lambda_k \rightarrow \infty$$

or equivalently,

$$\sup_{0 \leq t \leq T} \frac{1}{\sqrt{\lambda_k}} \left| \frac{w_I(1-\alpha)}{w_I(1-\alpha) + w_O\alpha} Q^{\lambda_k}(t) - Q_O^{\lambda_k}(t) \right| \rightarrow 0, \text{ in probability as } \lambda_k \rightarrow \infty.$$

It now follows that $\tilde{\epsilon}^{\lambda_k} \Rightarrow 0$ as $\lambda_k \rightarrow \infty$. Since the subsequence λ_k was arbitrary, we conclude that $\tilde{\epsilon}^\lambda \Rightarrow 0$ as $\lambda \rightarrow \infty$. \square

Proof of Corollary 1

The fact that $\frac{Q^\lambda(t)}{\sqrt{\lambda}} = \frac{Q_I^\lambda(t)}{\sqrt{\lambda}} + \frac{Q_O^\lambda(t)}{\sqrt{\lambda}}$ combined with Theorem 1 part (i) gives that for any $T > 0$,

$$\sup_{0 \leq t \leq T} \frac{1}{\sqrt{\lambda}} \left| \frac{w_I}{\alpha} \left(Q^\lambda(t) - Q_O^\lambda(t) \right) - \frac{w_O}{1-\alpha} Q_O^\lambda(t) \right| \rightarrow 0, \text{ in probability as } \lambda \rightarrow \infty$$

or equivalently,

$$\sup_{0 \leq t \leq T} \frac{1}{\sqrt{\lambda}} \left| \frac{w_I(1-\alpha)}{w_I(1-\alpha) + w_O\alpha} Q^\lambda(t) - Q_O^\lambda(t) \right| \rightarrow 0, \text{ in probability as } \lambda \rightarrow \infty.$$

But by Theorem 1 part (ii), we have that $\frac{1}{\sqrt{\lambda}} Q^\lambda \Rightarrow \tilde{Q}$, as $\lambda \rightarrow \infty$ and so Slutsky's theorem implies that

$$\frac{Q_O^\lambda}{\sqrt{\lambda}} \Rightarrow \frac{w_I(1-\alpha)}{w_I(1-\alpha) + w_O\alpha} \tilde{Q}, \text{ as } \lambda \rightarrow \infty. \quad (32)$$

Following a similar argument, we can conclude that

$$\frac{Q_I^\lambda}{\sqrt{\lambda}} \Rightarrow \frac{w_O\alpha}{w_I(1-\alpha) + w_O\alpha} \tilde{Q}, \text{ as } \lambda \rightarrow \infty.$$

For the last convergence result, first observe that for any $t \geq 0$

$$\frac{1}{\sqrt{\lambda}} N \left(\int_0^t \gamma [Q_O^\lambda(s) - 1]^+ ds \right) = \tilde{N}^\lambda (\bar{\tau}^\lambda(t)) + \gamma \int_0^t \frac{[Q_O^\lambda(s) - 1]^+}{\sqrt{\lambda}} ds.$$

By (18), $\tilde{N}^\lambda \circ \bar{\tau}^\lambda \Rightarrow 0$ as $\lambda \rightarrow \infty$. It now follows from (32) above that

$$\frac{1}{\sqrt{\lambda}} N \left(\int_0^\cdot \gamma [Q_O^\lambda(s) - 1]^+ ds \right) \Rightarrow \gamma \frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} \int_0^\cdot \tilde{Q}(s) ds \text{ as } \lambda \rightarrow \infty.$$

□

Proof of Theorem 2

Recalling that $\mathcal{W}_I(t) = Q_I(t)/[\mu(\lambda)\alpha]$, it then follows from Corollary 1 that

$$\tilde{\mathcal{W}}_I^\lambda(t) \Rightarrow \frac{w_O}{(1-\alpha)w_I + \alpha w_O} \tilde{Q}.$$

Hence it also follows from Lemma 3 that

$$\sup_{0 \leq t \leq T} \left| \tilde{W}_I^\lambda(t) - \tilde{\mathcal{W}}_I^\lambda(t) \right| \rightarrow 0 \text{ in probability as } \lambda \rightarrow \infty,$$

A very similar argument shows that for any $T > 0$

$$\sup_{0 \leq t \leq T} \left| \tilde{W}_O^\lambda(t) - \tilde{\mathcal{W}}_O^\lambda(t) \right| \rightarrow 0 \text{ in probability as } \lambda \rightarrow \infty,$$

except that since $\mathcal{W}_O(t) = O(t)/[\mu(\lambda)(1-\alpha)]$, recalling that $O(t)$ is upper-bounded in (10), we must additionally establish that

$$\frac{1}{\sqrt{\lambda}} \left(N \left(\int_0^t \gamma [Q_O(s) - 1]^+ ds \right) - N \left(\int_0^{[t - \sup_{0 \leq s \leq T} W_O(s)]^+} \gamma [Q_O(s) - 1]^+ ds \right) \right) \Rightarrow 0. \quad (33)$$

Note that

$$\begin{aligned} & \frac{1}{\sqrt{\lambda}} \left(N \left(\int_0^t \gamma [Q_O(s) - 1]^+ ds \right) - N \left(\int_0^{[t - \sup_{0 \leq s \leq T} W_O(s)]^+} \gamma [Q_O(s) - 1]^+ ds \right) \right) \\ &= \tilde{N}^\lambda(\bar{\tau}^\lambda(t)) - \tilde{N}^\lambda \left(\bar{\tau}^\lambda \left(\left[t - \sup_{0 \leq s \leq t} W_O^\lambda(s) \right]^+ \right) \right) + \gamma \int_{[t - \sup_{0 \leq s \leq t} W_O^\lambda(s)]^+}^t \frac{[Q_O^\lambda(s) - 1]^+}{\sqrt{\lambda}} ds. \end{aligned}$$

The first two terms in the above weakly converge to the zero process by the weak convergence in (18). Next, since the definition of the workload process P_O^λ implies $W_O^\lambda(t) \leq \frac{P_O^\lambda(t)}{1-\alpha}$ for all $t \geq 0$, it follows from Lemma 1 that $W_O^\lambda \rightarrow 0$ a.s., u.o.c., as $\lambda \rightarrow \infty$. Hence the third term also weakly converges to the zero process by the continuous mapping theorem and the weak

convergence of $Q^\lambda/\sqrt{\lambda}$ in Corollary 1. □

Proof of Proposition 1

We must show the following.

- (i) For any $T > 0$, $\sup_{0 \leq t \leq T} \left| \frac{w_I}{\alpha} \tilde{Q}_I^\lambda(t) - \frac{w_O}{1-\alpha} \tilde{Q}_O^\lambda(t) \right| \rightarrow 0$, in probability, as $\lambda \rightarrow \infty$.
- (ii) As $\lambda \rightarrow \infty$, $(\tilde{Q}^\lambda, \tilde{I}^\lambda) \Rightarrow (\tilde{Q}, \tilde{I})$.
- (iii) As $\lambda \rightarrow \infty$,

$$\sup_{0 \leq t \leq T} \sqrt{\lambda} |W_I(t) - \mathcal{W}_I(t)| \rightarrow 0 \text{ and } \sup_{0 \leq t \leq T} \sqrt{\lambda} |W_O(t) - \mathcal{W}_O(t)| \rightarrow 0, \text{ in probability.}$$

Proof of (i)

Modify the definitions of \tilde{U}_1^λ and \tilde{U}_2^λ in the proof of Theorem 1(i) so that

$$\begin{aligned} \tilde{U}_1^\lambda(t, s) &= -\frac{w_O}{1-\alpha} \left(\tilde{A}^\lambda(t) - \tilde{A}^\lambda(s) \right) \\ &\quad + \left\{ \frac{\theta}{\sqrt{\lambda}} (w_I - w_O) - \frac{\alpha}{1-\alpha} w_O - w_I + \frac{w_O}{1-\alpha} \frac{1}{\lambda l^\lambda} \right\} \sqrt{\lambda} (t - s) \\ \tilde{U}_2^\lambda(t, s) &= -\frac{w_I}{\alpha} \left(\tilde{A}^\lambda(t) - \tilde{A}^\lambda(s) \right) \\ &\quad + \left\{ \frac{\theta}{\sqrt{\lambda}} (w_O - w_I) - \frac{1-\alpha}{\alpha} w_I - w_O + \frac{w_I}{\alpha} \frac{1}{\lambda l^\lambda} \right\} \sqrt{\lambda} (t - s). \end{aligned}$$

With ξ_λ and ξ_λ^* defined exactly as in the proof of Theorem 1(i), observe that when

$$\frac{w_I}{\alpha} \tilde{Q}_I^\lambda(\xi_\lambda^*) > (\leq) \frac{w_O}{1-\alpha} \tilde{Q}_O^\lambda(\xi_\lambda^*),$$

because the inline (offline) queue does not become empty during $[\xi_\lambda^*, \xi_\lambda]$, the offline (inline) queue may become empty, and service occurs in discrete time intervals

$$\begin{aligned} S_I^\lambda(t) - S_I^\lambda(\xi_\lambda^* -) &\geq (\leq) \left\lfloor \frac{t - \xi_\lambda^*}{l^\lambda} \right\rfloor \lfloor \alpha n^\lambda \rfloor \\ S_O^\lambda(t) - S_O^\lambda(\xi_\lambda^* -) &\leq (\geq) \left\lceil \frac{t - \xi_\lambda^*}{l^\lambda} \right\rceil \lceil (1-\alpha) n^\lambda \rceil. \end{aligned}$$

Then, substitution of the above bounds into the equivalent of (24) in the proof of Theorem 1(i) in this setting (specifically, replace $S_I^\lambda(T_I^\lambda(t)) - S_I^\lambda(T_I^\lambda(\xi_\lambda^* -))$ with $S_I^\lambda(t) - S_I^\lambda(\xi_\lambda^* -)$)

and $S_O^\lambda(T_O^\lambda(t)) - S_O^\lambda(T_O^\lambda(\xi_\lambda^*-))$ with $S_O^\lambda(t) - S_O^\lambda(\xi_\lambda^*-)$ shows that for large enough λ

$$\left| \frac{w_I}{\alpha} \tilde{Q}_I^\lambda(t) - \frac{w_O}{1-\alpha} \tilde{Q}_O^\lambda(t) \right| \leq \frac{\epsilon}{2} + 1 + \max \left\{ \tilde{U}_1^\lambda(t, \xi_\lambda^*-), \tilde{U}_2^\lambda(t, \xi_\lambda^*-) \right\} + \frac{w_O}{1-\alpha} \frac{1}{\sqrt{\lambda}} N \left(\int_0^t \gamma Q_O^\lambda(s) ds \right).$$

Since $(\lambda l^\lambda)^{-1} = \lambda^{-1/3} \rightarrow 0$ as $\lambda \rightarrow \infty$, the remainder of the proof proceeds exactly as the proof of Theorem 1 (i), noting that by Lemma 4

$$\frac{1}{\lambda} \int_0^t \gamma Q_O^\lambda(s) ds \rightarrow 0 \text{ a.s., u.o.c.,}$$

as $\lambda \rightarrow \infty$.

Proof of (ii)

We first obtain a useful equivalent representation for the process $Q^\lambda(t) = Q_I^\lambda(t) + Q_O^\lambda(t)$. Define

$$\epsilon^\lambda(t) \equiv \gamma \int_0^t \left(\frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} Q^\lambda(s) - Q_O^\lambda(s) \right) ds,$$

so that

$$\begin{aligned} Q^\lambda(t) &= A^\lambda(t) - S_I^\lambda(t) - S_O^\lambda(t) - \int_0^t \gamma \frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} Q^\lambda(s) ds \\ &\quad + \epsilon^\lambda(t) - N \left(\int_0^t \gamma Q_O^\lambda(s) ds \right) + \int_0^t \gamma Q_O^\lambda(s) ds. \end{aligned}$$

Next we define recursively the process that tracks the cumulative number of empty ride seats

$$\begin{aligned} I^\lambda(0) &\equiv 0 \\ I^\lambda(il^\lambda) &\equiv I^\lambda((i-1)l^\lambda) + [n^\lambda - Q^\lambda(il^\lambda-)]^+. \end{aligned}$$

The process I^λ does not increase in between the discrete time points $l^\lambda, 2l^\lambda, 3l^\lambda, \dots$, and so for any $t > 0$,

$$I^\lambda(t) = I^\lambda \left(\left\lfloor \frac{t}{l^\lambda} \right\rfloor l^\lambda \right).$$

Then,

$$\begin{aligned} S_I^\lambda(t) + S_O^\lambda(t) &= \sum_{i=1}^{\lfloor t/l^\lambda \rfloor} \left(B_I^\lambda(il^\lambda) + B_O^\lambda(il^\lambda) \right) \\ &= \sum_{i=1}^{\lfloor t/l^\lambda \rfloor} \left(n^\lambda \mathbf{1}\{Q^\lambda(il^\lambda) \geq n^\lambda\} + Q^\lambda(il^\lambda) \mathbf{1}\{Q^\lambda(il^\lambda) < n^\lambda\} \right), \end{aligned}$$

and so

$$S_I^\lambda(t) + S_O^\lambda(t) + I^\lambda(t) = \left\lfloor \frac{t}{l^\lambda} \right\rfloor n^\lambda.$$

It follows that

$$Q^\lambda(t) = X^\lambda(t) + \epsilon^\lambda(t) - \int_0^t \gamma \frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} Q^\lambda(s) ds + I^\lambda(t) \quad (34)$$

where

$$X^\lambda(t) \equiv A^\lambda(t) - \left\lfloor \frac{t}{l^\lambda} \right\rfloor n^\lambda - N \left(\int_0^t \gamma Q_O^\lambda(s) ds \right) + \int_0^t \gamma Q_O^\lambda(s) ds.$$

Furthermore, because the ride will depart with empty seats only when no customers are waiting,

$$\sum_{i=1}^{\infty} Q^\lambda(il^\lambda) (I^\lambda(il^\lambda) - I^\lambda((i-1)l^\lambda)) = 0. \quad (35)$$

By Lemma 4, there exists a convergent subsequence

$$\left(\frac{Q^{\lambda_i}}{\sqrt{\lambda_i}}, \frac{I^{\lambda_i}}{\sqrt{\lambda_i}} \right) \Rightarrow (\tilde{Q}, \tilde{I}) \text{ as } \lambda_i \rightarrow \infty.$$

We show that the limit (\tilde{Q}, \tilde{I}) satisfies

$$(\tilde{Q}, \tilde{I}) = (\phi^\kappa, \psi^\kappa) (\tilde{X}), \quad (36)$$

where \tilde{X} is a Brownian motion with drift θ and variance σ_A^2 . We first show (C1) in Definition 1 is satisfied. On the subsequence λ_i , because

$$\frac{X^{\lambda_i}(t)}{\sqrt{\lambda_i}} = \tilde{A}^{\lambda_i}(t) - \tilde{N}^{\lambda_i}(\tilde{\tau}^{\lambda_i}(t)) + \sqrt{\lambda_i} t \left(1 - \left\lfloor \frac{t}{l^{\lambda_i}} \right\rfloor \frac{l^{\lambda_i}}{t} \right) + \theta t \left\lfloor \frac{t}{l^{\lambda_i}} \right\rfloor \frac{l^{\lambda_i}}{t},$$

the functional central limit theorem shows

$$\frac{X^{\lambda_i}}{\sqrt{\lambda_i}} \Rightarrow \tilde{X}.$$

Also on the subsequence λ_i , as in the proof of part (ii) of Theorem 1 (which requires the tightness of \tilde{Q}^λ established in Lemma 4),

$$\frac{\epsilon_i^\lambda}{\sqrt{\lambda_i}} \Rightarrow 0,$$

as $\lambda_i \rightarrow \infty$. We conclude from (34) that the limit (\tilde{Q}, \tilde{I}) satisfies

$$\tilde{Q}(t) = \tilde{X}(t) - \int_0^t \gamma \frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} \tilde{Q}(s) ds + \tilde{I}(t). \quad (37)$$

For (C2), note that on the subsequence λ_i , the definition of the Reimann-Stieltjes integral implies that when we take limits as $\lambda_i \rightarrow \infty$ on both sides of the equality in (35),

$$\int_0^\infty \tilde{Q}(t) d\tilde{I}(t) = 0.$$

Since furthermore $I^{\lambda_i}(0) = 0$ and I^{λ_i} is non-decreasing, it follows that $\tilde{I}(0) = 0$ and \tilde{I} is non-decreasing. We conclude that (36) is valid. From the representation (22), this is equivalent to the stochastic equation for \tilde{Q} in (11). Since the subsequence λ_i was arbitrary, this part of the proof is complete.

Proof of (iii)

The argument is very similar to the proof of Theorem 2, and so is omitted. The exception is that the upper-bound on O^λ in (17) replaces the upper-bound on O^λ in (10).

□

Proofs of Lemmas 1 - 4

The proofs of Lemmas 1-4 require use of the well-known conventional one-sided regulator mapping. This mapping is defined exactly as in Definition 1 for $\kappa = 0$.

Proof of Lemma 1

We first represent the process \bar{Q}^λ using the conventional one-sided regulator mapping. Define

$$\bar{X}^\lambda(t) \equiv \bar{A}^\lambda(t) - \bar{S}_I^\lambda(T_I^\lambda(t)) - \bar{S}_O^\lambda(T_O^\lambda(t)) - \bar{N}^\lambda(\bar{\tau}^\lambda(t)) + \frac{\theta}{\sqrt{\lambda}}t.$$

Then, for all $t \geq 0$,

$$\bar{Q}^\lambda(t) = \bar{X}^\lambda(t) - \bar{\tau}^\lambda(t) + \left(1 - \frac{\theta}{\sqrt{\lambda}}\right) I^\lambda(t).$$

Since I^λ is non-decreasing, $I^\lambda(0) = 0$ and $\int_0^\infty \bar{Q}^\lambda(t) d\left(\left(1 - \frac{\theta}{\sqrt{\lambda}}\right) I^\lambda(t)\right) = 0$, it follows that

$$\left(\bar{Q}^\lambda, \left(1 - \frac{\theta}{\sqrt{\lambda}}\right) I^\lambda\right) = (\phi, \psi) \left(\bar{X}^\lambda - \bar{\tau}^\lambda\right).$$

Since $\bar{\tau}^\lambda$ is a non-decreasing process, Lemma 5.1 in Kruk et al. (2007) establishes

$$\phi \left(\bar{X}^\lambda - \bar{\tau}^\lambda\right) \leq \phi \left(\bar{X}^\lambda\right).$$

The functional strong law of large numbers establishes

$$\bar{X}^\lambda \rightarrow 0 \text{ a.s., u.o.c.,}$$

as $\lambda \rightarrow \infty$, and so, by the continuous mapping theorem,

$$\phi \left(\bar{X}^\lambda\right) \rightarrow 0 \text{ a.s., u.o.c..}$$

Since \bar{Q}^λ is a non-negative process bounded above by $\phi \left(\bar{X}^\lambda\right)$, we conclude

$$\bar{Q}^\lambda \rightarrow 0 \text{ a.s., u.o.c.,}$$

as $\lambda \rightarrow \infty$, and so also

$$\bar{\tau}^\lambda \rightarrow 0 \text{ a.s., u.o.c.,}$$

as $\lambda \rightarrow \infty$. Since $(\phi, \psi)(0) = (0, 0)$ and ψ is a continuous function in D , we can also conclude that

$$I^\lambda = \frac{\sqrt{\lambda}}{\sqrt{\lambda} - \theta} \psi \left(\bar{X}^\lambda - \bar{\tau}^\lambda\right) \rightarrow 0 \text{ a.s., u.o.c.,}$$

as $\lambda \rightarrow \infty$. The condition (8) then implies

$$T_I^\lambda + T_O^\lambda \rightarrow e \text{ a.s., u.o.c.,}$$

as $\lambda \rightarrow \infty$.

It remains to show

$$P_O^\lambda \rightarrow 0 \text{ and } P_I^\lambda \rightarrow 0 \text{ a.s., u.o.c.,}$$

as $\lambda \rightarrow \infty$. Define the total time required to process all the customers in the offline queue, ignoring any partial processing that may have already occurred on the customer in service

$$U_O^\lambda(t) \equiv \sum_{j=S_O^\lambda(T_O^\lambda(t))+1}^{S_O^\lambda(T_O^\lambda(t))+Q_O^\lambda(t)} \frac{v_j^O}{\mu(\lambda)}.$$

Then,

$$P_O^\lambda(t) \leq U_O^\lambda(t) \text{ for all } t \geq 0.$$

Define

$$\bar{V}_O^\lambda(t) \equiv \frac{1}{\lambda} \sum_{i=1}^{\lfloor \lambda t \rfloor} (v_i^O - 1),$$

and observe that

$$U_O^\lambda(t) = \frac{\sqrt{\lambda}}{\sqrt{\lambda} - \theta} \left(\bar{V}_O^\lambda \left(\frac{1}{\lambda} S_O^\lambda(T_O^\lambda(t)) + \bar{Q}_O^\lambda(t) \right) - \bar{V}_O^\lambda \left(\frac{1}{\lambda} S_O^\lambda(T_O^\lambda(t)) \right) \right) + \frac{\sqrt{\lambda}}{\sqrt{\lambda} - \theta} \bar{Q}_O^\lambda(t).$$

Since $0 \leq \bar{Q}_O^\lambda(t) \leq \bar{Q}^\lambda(t)$ for all $t \geq 0$ and we have already established $\bar{Q}^\lambda \rightarrow 0$ a.s., u.o.c. as $\lambda \rightarrow \infty$, it follows that

$$\bar{Q}_O^\lambda \rightarrow 0 \text{ a.s., u.o.c.,}$$

as $\lambda \rightarrow \infty$. Therefore, because also $\bar{V}_O^\lambda \rightarrow 0$ a.s., u.o.c. as $\lambda \rightarrow \infty$, it follows that $\bar{U}_O^\lambda \rightarrow 0$ a.s., u.o.c. as $\lambda \rightarrow \infty$. We conclude

$$P_O^\lambda \rightarrow 0 \text{ a.s., u.o.c.,}$$

as $\lambda \rightarrow \infty$. The argument that $P_I^\lambda \rightarrow 0$ a.s., u.o.c. is identical and so is omitted. \square

Proof of Lemma 2

The representation (29), and the continuity of the mapping $(\phi^\kappa, \psi^\kappa)$ established in Proposition 4.1 part (iv) in Reed and Ward imply that it is enough to show that the families $\{\tilde{X}^\lambda\}$ and $\{\tilde{\epsilon}^\lambda\}$ defined in the proof of part (ii) of Theorem 1 are tight in D . The tightness of the family $\{\tilde{X}^\lambda\}$ is immediate from the weak convergence established in (30), which requires only Lemma 1. Hence we need only show that the family $\{\tilde{\epsilon}^\lambda\}$ is tight in D . (Note that this does not follow from the weak convergence in (31) because that argument relies on the fact that the sequence $\{\tilde{Q}^\lambda\}$ is tight.) For this, we must verify conditions (16.17) and (16.18) in Billingsley. Suppose we can show that for any $T > 0$ and $\epsilon > 0$ arbitrarily small, there exists

B and λ_0 large enough such that

$$P\left(\sup_{0 \leq t \leq T} \tilde{Q}^\lambda(t) > B\right) < \epsilon \quad (38)$$

for all $\lambda \geq \lambda_0$.

(B16.17) We must show that for $\eta > 0$ arbitrarily small, there exists an a and a λ_0 large enough such that

$$P\left(\sup_{0 \leq t \leq T} |\tilde{\epsilon}^\lambda(t)| \geq a\right) < \eta, \quad \lambda \geq \lambda_0.$$

Since

$$\tilde{\epsilon}^\lambda(t) \leq \gamma T \frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} \sup_{0 \leq t \leq T} \tilde{Q}^\lambda(t),$$

this follows from (38).

(B16.18) It is sufficient to show that for $\gamma > 0$ and $\eta > 0$ arbitrarily small, there exists a δ small enough and a λ_0 large enough such that

$$P\left(\sup_{0 \leq t \leq T-\delta} \sup_{v, s \in [t, t+\delta]} |\tilde{\epsilon}^\lambda(s) - \tilde{\epsilon}^\lambda(v)| \geq \gamma\right) < \eta, \quad \lambda \geq \lambda_0.$$

Since

$$|\tilde{\epsilon}^\lambda(s) - \tilde{\epsilon}^\lambda(v)| \leq \gamma \int_v^s \left| \frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} \tilde{Q}^\lambda(\zeta) \right| d\zeta + \gamma \int_v^s \tilde{Q}_O^\lambda(\zeta) d\zeta$$

and $\tilde{Q}_O^\lambda(t) \leq \tilde{Q}^\lambda(t)$ for all $t \geq 0$,

$$\sup_{0 \leq t \leq T-\delta} \sup_{v, s \in [t, t+\delta]} |\tilde{\epsilon}^\lambda(s) - \tilde{\epsilon}^\lambda(v)| \leq \gamma \left(\frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} + 1 \right) \delta \sup_{0 \leq t \leq T} \tilde{Q}^\lambda(t).$$

Hence the condition (B16.18) also follows from (38).

Finally, it remains to establish (38). Fix $T > 0$ and $\epsilon > 0$. We first represent the process \tilde{Q}^λ using the conventional one-sided regulator mapping. Define

$$\begin{aligned} \tilde{\chi}^\lambda &\equiv \tilde{A}^\lambda(t) - \tilde{S}_I^\lambda(T_I^\lambda(t)) - \tilde{S}_O^\lambda(T_O^\lambda(t)) + \theta t \\ \tilde{\mathcal{A}}^\lambda(t) &\equiv \frac{1}{\sqrt{\lambda}} N \left(\int_0^t \gamma [Q_O^\lambda(s) - 1]^+ ds \right). \end{aligned}$$

Then,

$$\tilde{Q}^\lambda(t) = \tilde{\chi}^\lambda(t) - \tilde{\mathcal{A}}^\lambda(t) + \tilde{I}^\lambda(t) \geq 0 \text{ for all } t \geq 0.$$

Since \tilde{I}^λ is non-decreasing, $\tilde{I}^\lambda(0) = 0$, it follows that

$$\left(\tilde{Q}^\lambda, \tilde{I}^\lambda\right) = (\phi, \psi) \left(\tilde{\chi}^\lambda - \tilde{\mathcal{A}}^\lambda\right). \quad (39)$$

Since $\tilde{\mathcal{A}}^\lambda$ is a non-decreasing process, Lemma 5.1 in Kruk et al. (2007) establishes that

$$\phi \left(\tilde{\chi}^\lambda - \tilde{\mathcal{A}}^\lambda\right) (t) \leq \phi \left(\tilde{\chi}^\lambda\right) (t) \text{ for all } t \geq 0. \quad (40)$$

The functional central limit theorem and the continuous mapping theorem establish

$$\phi \left(\tilde{\chi}^\lambda\right) \Rightarrow \phi \left(\tilde{X}\right),$$

as $\lambda \rightarrow \infty$, where \tilde{X} is a Brownian motion with drift θ and variance $\sigma^2 = \sigma_A^2 + \sigma_S^2$ as in (11). Since weak convergence implies the random variable $\sup_{0 \leq t \leq T} \phi \left(\tilde{\chi}^\lambda\right) (t)$ is tight, there exists B and λ_0 large enough so that for all $\lambda \geq \lambda_0$

$$P \left(\sup_{0 \leq t \leq T} \phi \left(\tilde{\chi}^\lambda\right) (t) > B \right) < \epsilon.$$

Therefore, it follows from the representation (39) and the upper bound (40) that for all $\lambda \geq \lambda_0$

$$P \left(\sup_{0 \leq t \leq T} \tilde{Q}^\lambda(t) > B \right) < \epsilon.$$

□

Proof of Lemma 3

An argument very similar to Theorem 5.3 in Reiman (1984) shows that

$$\tilde{P}_I^\lambda \Rightarrow \frac{\alpha w_O}{(1 - \alpha)w_I + \alpha w_O} \tilde{Q} \text{ as } \lambda \rightarrow \infty.$$

Suppose we can also show that

$$\tilde{P}_O^\lambda \Rightarrow \frac{(1 - \alpha)w_I}{(1 - \alpha)w_I + \alpha w_O} \tilde{Q} \text{ as } \lambda \rightarrow \infty. \quad (41)$$

Next note that it follows from Corollary 1 that for any $t > 0$

$$P \left(Q_I^\lambda(t) > 0\right) \rightarrow 1 \text{ and } P \left(Q_O^\lambda(t) > 0\right) \rightarrow 1,$$

as $\lambda \rightarrow \infty$. Since $\frac{d}{dt}T_I^\lambda(t) = \alpha$ and $\frac{d}{dt}T_O^\lambda(t) = (1 - \alpha)$ if and only if $Q_I^\lambda(t) > 0$ and $Q_O^\lambda(t) > 0$, we conclude

$$P\left(\frac{d}{dt}T_I^\lambda(t) = \alpha\right) \rightarrow 1 \text{ and } P\left(\frac{d}{dt}T_O^\lambda(t) = 1 - \alpha\right) \rightarrow 1,$$

as $\lambda \rightarrow \infty$. The convergences in (19) and (20) now follow by the definition of the workload process and the converging together Lemma.

It remains to show (41). Since the offline service queue receives at least $(1 - \alpha)$ proportion of the server's effort when the queue is non-empty, $(1 - \alpha)^{-1}P_O^\lambda(t)$ upper bounds the amount of time required to finish serving all customers in the offline queue that will eventually receive service. Therefore, at time $t > 0$, the number of customers in the offline queue that will eventually abandon is less than or equal to

$$\mathcal{A}^\lambda(t) \equiv N\left(\int_0^{t+(1-\alpha)^{-1}P_O^\lambda(t)} \gamma[Q_O^\lambda(s) - 1]^+ ds\right) - N\left(\int_0^t \gamma[Q_O^\lambda(s) - 1]^+ ds\right).$$

Then, $Q_O^\lambda(t) - \mathcal{A}^\lambda(t)$ is a lower bound on the number of customers in the offline queue that will eventually receive service, and so

$$L_O^\lambda(t) \equiv \sum_{j=S_O^\lambda(T_O^\lambda(t))+2}^{S_O^\lambda(T_O^\lambda(t))+Q_O^\lambda(t)-\mathcal{A}^\lambda(t)} \frac{v_j^O}{\mu(\lambda)} \leq P_O^\lambda(t).$$

Also, $Q_O^\lambda(t)$ is an upper bound on the number of customers in the offline queue that will eventually receive service, and so

$$U_O^\lambda(t) \equiv \sum_{j=S_O^\lambda(T_O^\lambda(t))+1}^{S_O^\lambda(T_O^\lambda(t))+Q_O^\lambda(t)} \frac{v_j^O}{\mu(\lambda)} \geq P_O^\lambda(t).$$

Note that to get the upper bound of the workload process we include in the summation the customer in service, whereas to get the lower bound, we do not. We conclude

$$0 \leq \sqrt{\lambda}P_O^\lambda(t) - \sqrt{\lambda}L_O^\lambda(t) \leq \sqrt{\lambda}U_O^\lambda(t) - \sqrt{\lambda}L_O^\lambda(t). \quad (42)$$

Define

$$\tilde{V}_O^\lambda(t) \equiv \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{\lfloor \lambda t \rfloor} (v_i^O - 1) \text{ for all } t \geq 0.$$

Observe that

$$\begin{aligned}
& \sqrt{\lambda}U_{\mathcal{O}}^{\lambda}(t) - \sqrt{\lambda}L_{\mathcal{O}}^{\lambda}(t) \\
&= \frac{\sqrt{\lambda}}{\mu(\lambda)}v_{S_{\mathcal{O}}^{\lambda}(T_{\mathcal{O}}^{\lambda}(t))+1}}^{\mathcal{O}} + \frac{\sqrt{\lambda}}{\mu(\lambda)}\mathcal{A}^{\lambda}(t) \\
&\quad + \frac{\lambda}{\mu(\lambda)}\left(\tilde{V}_{\mathcal{O}}^{\lambda}\left(\frac{S_{\mathcal{O}}^{\lambda}(T_{\mathcal{O}}^{\lambda}(t))}{\lambda} + \frac{Q_{\mathcal{O}}^{\lambda}(t)}{\lambda}\right) - \tilde{V}_{\mathcal{O}}^{\lambda}\left(\frac{S_{\mathcal{O}}^{\lambda}(T_{\mathcal{O}}^{\lambda}(t))}{\lambda} + \frac{Q_{\mathcal{O}}^{\lambda}(t)}{\lambda} - \frac{\mathcal{A}^{\lambda}(t)}{\lambda}\right)\right)
\end{aligned} \tag{43}$$

and

$$\begin{aligned}
& \sqrt{\lambda}L_{\mathcal{O}}^{\lambda}(t) \\
&= \frac{\lambda}{\mu(\lambda)}\tilde{Q}_{\mathcal{O}}^{\lambda}(t) - \frac{\sqrt{\lambda}}{\mu(\lambda)} - \frac{\sqrt{\lambda}}{\mu(\lambda)}\mathcal{A}^{\lambda}(t) \\
&\quad + \frac{\lambda}{\mu(\lambda)}\left(\tilde{V}_{\mathcal{O}}^{\lambda}\left(\frac{S_{\mathcal{O}}^{\lambda}(T_{\mathcal{O}}^{\lambda}(t))}{\lambda} + \frac{Q_{\mathcal{O}}^{\lambda}(t)}{\lambda} - \frac{\mathcal{A}^{\lambda}(t)}{\lambda}\right) - \tilde{V}_{\mathcal{O}}^{\lambda}\left(\frac{S_{\mathcal{O}}^{\lambda}(T_{\mathcal{O}}^{\lambda}(t))}{\lambda} + \frac{1}{\lambda}\right)\right).
\end{aligned} \tag{44}$$

We will first show that $\sqrt{\lambda}U_{\mathcal{O}}^{\lambda} - \sqrt{\lambda}L_{\mathcal{O}}^{\lambda} \Rightarrow 0$ as $\lambda \rightarrow \infty$, and then show

$$\sqrt{\lambda}L_{\mathcal{O}}^{\lambda} \Rightarrow \frac{(1-\alpha)w_I}{(1-\alpha)w_I + \alpha w_{\mathcal{O}}}\tilde{Q} \tag{45}$$

as $\lambda \rightarrow \infty$. The inequality (42) and the converging together lemma then establish (41).

To show $\sqrt{\lambda}U_{\mathcal{O}}^{\lambda} - \sqrt{\lambda}L_{\mathcal{O}}^{\lambda} \Rightarrow 0$ as $\lambda \rightarrow \infty$, first note that it follows from Lemma 3 in Iglehart and Whitt (1970) that for any $t > 0$

$$\frac{\sqrt{\lambda}}{\mu(\lambda)}v_{S_{\mathcal{O}}^{\lambda}(T_{\mathcal{O}}^{\lambda}(t))+1}}^{\mathcal{O}} = \frac{\lambda}{\lambda - \sqrt{\lambda}\theta} \frac{1}{\sqrt{\lambda}}v_{S_{\mathcal{O}}^{\lambda}(T_{\mathcal{O}}^{\lambda}(t))+1}}^{\mathcal{O}} \rightarrow 0$$

in probability, as $\lambda \rightarrow \infty$. Next, since

$$\frac{1}{\sqrt{\lambda}}\mathcal{A}^{\lambda}(t) = \tilde{N}^{\lambda}\left(\bar{\tau}^{\lambda}\left(t + \frac{P_{\mathcal{O}}^{\lambda}(t)}{1-\alpha}\right)\right) - \tilde{N}^{\lambda}(\bar{\tau}^{\lambda}(t)) + \gamma \int_t^{t+(1-\alpha)^{-1}P_{\mathcal{O}}^{\lambda}(t)} \frac{[Q_{\mathcal{O}}^{\lambda}(s) - 1]^+}{\sqrt{\lambda}} ds,$$

and Lemma 1 establishes $\bar{\tau}^{\lambda} \rightarrow 0$ and $P_{\mathcal{O}}^{\lambda} \rightarrow 0$ a.s., u.o.c., it follows from the functional central limit theorem, continuous mapping theorem, and the weak convergence of $\frac{Q_{\mathcal{O}}^{\lambda}}{\sqrt{\lambda}}$ in Corollary 1 that $\lambda^{-1/2}\mathcal{A}^{\lambda} \Rightarrow 0$ and so

$$\frac{\sqrt{\lambda}}{\mu(\lambda)}\mathcal{A}^{\lambda} = \frac{\lambda}{\lambda - \sqrt{\lambda}\theta} \frac{1}{\sqrt{\lambda}}\mathcal{A}^{\lambda} \Rightarrow 0 \tag{46}$$

as $\lambda \rightarrow \infty$. Now, the sequence $\{T_{\mathcal{O}}^{\lambda}\}$ is tight in D because $|T_{\mathcal{O}}^{\lambda}(t) - T_{\mathcal{O}}^{\lambda}(s)| \leq |t - s|$. On any

subsequence $\{\lambda_k\}$ on which

$$T_O^{\lambda_k} \Rightarrow T_O$$

as $\lambda_k \rightarrow \infty$, the functional strong law of large numbers and random time change theorem establish

$$\frac{S_O^{\lambda_k} \circ T_O^{\lambda_k}}{\lambda_k} \Rightarrow T_O$$

as $\lambda_k \rightarrow \infty$. Furthermore, on this same subsequence, by the convergences in (46) and Lemma 1, $\lambda_k^{-1} \mathcal{A}^{\lambda_k} \Rightarrow 0$ and $\lambda_k^{-1} Q_O^{\lambda_k} \rightarrow 0$ a.s., u.o.c. as $\lambda_k \rightarrow \infty$. Therefore, because by Donsker's theorem \tilde{V}_O^λ weakly converges to a continuous limit process,

$$\tilde{V}_O^{\lambda_k} \left(\frac{S_O^{\lambda_k}(T_O^{\lambda_k}(\cdot))}{\lambda_k} + \frac{Q_O^{\lambda_k}(\cdot)}{\lambda_k} \right) - \tilde{V}_O^{\lambda_k} \left(\frac{S_O^{\lambda_k}(T_O^{\lambda_k}(\cdot))}{\lambda_k} + \frac{Q_O^{\lambda_k}(\cdot)}{\lambda_k} - \frac{\mathcal{A}^{\lambda_k}(\cdot)}{\lambda_k} \right) \Rightarrow 0$$

as $\lambda_k \rightarrow \infty$. Since the subsequence $\{\lambda_k\}$ was arbitrary, it follows that

$$\tilde{V}_O^\lambda \left(\frac{S_O^\lambda(T_O^\lambda(\cdot))}{\lambda} + \frac{Q_O^\lambda(\cdot)}{\lambda} \right) - \tilde{V}_O^\lambda \left(\frac{S_O^\lambda(T_O^\lambda(\cdot))}{\lambda} + \frac{Q_O^\lambda(\cdot)}{\lambda} - \frac{\mathcal{A}^\lambda(\cdot)}{\lambda} \right) \Rightarrow 0$$

as $\lambda \rightarrow \infty$. We conclude from (43) that as $\lambda \rightarrow \infty$

$$\sqrt{\lambda} U_O^\lambda - \sqrt{\lambda} L_O^\lambda \Rightarrow 0.$$

We now establish the weak convergence in (45). An argument similar to that in the above paragraph shows

$$\tilde{V}_O^\lambda \left(\frac{S_O^\lambda(T_O^\lambda(\cdot))}{\lambda} + \frac{Q_O^\lambda(\cdot)}{\lambda} - \frac{\mathcal{A}^\lambda(\cdot)}{\lambda} \right) - \tilde{V}_O^\lambda \left(\frac{S_O^\lambda(T_O^\lambda(\cdot))}{\lambda} + \frac{1}{\lambda} \right) \Rightarrow 0$$

as $\lambda \rightarrow \infty$. Hence, the representation of $\sqrt{\lambda} L_O^\lambda$ in (44), Corollary 1, the convergence in (46), and the continuous mapping theorem establish (45). □

Proof of Lemma 4

We divide the proof of Lemma 4 into three parts, with each part re-proving Lemmas 1, 2, and 3 for the modified model in Section 5, in which customers are served in batches at set time intervals.

Proof of (21) (Lemma 1 equivalent)

We require defining the following two comparison systems. Comparison system 1 is the model in Section 5 without abandonments. In particular, the inline and offline queue-length processes, $Q_{B,I}^\lambda$ and $Q_{B,O}^\lambda$, satisfy equations (15) and (16) with $\gamma = 0$. Under the same arrival sequence, on a sample path basis,

$$Q^\lambda(t) \leq Q_{B,I}^\lambda(t) + Q_{B,O}^\lambda(t), \text{ for all } t \geq 0.$$

Comparison system 2 is a conventional single-server queue with no abandonments and deterministic, non-batched service. In particular, the queue-length process evolution equation is

$$Q_C^\lambda(t) \equiv A^\lambda(t) - \mu(\lambda) \int_0^t \mathbf{1}\{Q_C^\lambda(s) > 0\} ds.$$

Under the same arrival sequence, on a sample path basis

$$Q_{B,I}^\lambda(il^\lambda) + Q_{B,O}^\lambda(il^\lambda) = Q_C^\lambda(il^\lambda) \text{ for every } i = 0, 1, 2, \dots$$

We conclude that on a sample path basis

$$Q^\lambda(t) \leq Q_C^\lambda\left(\left\lfloor \frac{t}{l^\lambda} \right\rfloor l^\lambda\right) + A^\lambda(t) - A^\lambda\left(\left\lfloor \frac{t}{l^\lambda} \right\rfloor l^\lambda\right). \quad (47)$$

It is well known that $Q_C^\lambda/\lambda \rightarrow 0$ a.s., u.o.c., as $\lambda \rightarrow \infty$. Furthermore,

$$\frac{A^\lambda(t) - A^\lambda\left(\left\lfloor \frac{t}{l^\lambda} \right\rfloor l^\lambda\right)}{\lambda} = \bar{A}^\lambda(t) - \bar{A}^\lambda\left(\left\lfloor \frac{t}{l^\lambda} \right\rfloor l^\lambda\right) + \left(t - \left\lfloor \frac{t}{l^\lambda} \right\rfloor l^\lambda\right),$$

and, as $\lambda \rightarrow \infty$, $\bar{A}^\lambda \rightarrow 0$ a.s., u.o.c. and $(t - \lfloor t/l^\lambda \rfloor l^\lambda) \rightarrow 0$. Hence

$$\bar{Q}^\lambda \rightarrow 0 \text{ a.s., u.o.c.,}$$

as $\lambda \rightarrow \infty$. It then follows that

$$\bar{\tau}^\lambda \rightarrow 0 \text{ a.s., u.o.c.,}$$

as $\lambda \rightarrow \infty$.

Proof that the sequence $\{\tilde{Q}^\lambda, \tilde{I}^\lambda\}$ is tight in D (Lemma 2 equivalent)

It is sufficient to verify that the sequence $\{\tilde{Q}^\lambda\}$ is tight in D . Tightness of the sequence $\{\tilde{Q}^\lambda, \tilde{I}^\lambda\}$ then follows from equation (34), because the functional central limit theorem shows the sequence $\{X^\lambda/\sqrt{\lambda}\}$ is tight, and tightness of the sequence $\{\epsilon^\lambda/\sqrt{\lambda}\}$ can be established

very similarly to Lemma 2. Let $T > 0$. We verify conditions (16.17) and (16.18) in Theorem 16.8 in Billingsley.

(B16.17) We must show that for $\eta > 0$ arbitrarily small, there exists an a and a λ_0 large enough such that

$$P \left(\sup_{0 \leq t \leq T} \left| \tilde{Q}^\lambda(t) \right| \geq a \right) < \eta, \quad \lambda \geq \lambda_0. \quad (48)$$

It is well-known that for Q_C^λ defined as in the first part of this proof, $Q_C^\lambda/\sqrt{\lambda}$ weakly converges to a reflected Brownian motion with drift θ and variance σ_A^2 . Furthermore,

$$\frac{A^\lambda(t) - A^\lambda \left(\lfloor \frac{t}{l^\lambda} \rfloor l^\lambda \right)}{\sqrt{\lambda}} = \tilde{A}^\lambda(t) - \tilde{A}^\lambda \left(\left\lfloor \frac{t}{l^\lambda} \right\rfloor l^\lambda \right) + \sqrt{\lambda} \left(t - \left\lfloor \frac{t}{l^\lambda} \right\rfloor l^\lambda \right).$$

The functional central limit theorem implies that \tilde{A}^λ weakly converges to a Brownian motion, and the definition of l^λ implies $\sqrt{\lambda} \left(t - \lfloor t/l^\lambda \rfloor l^\lambda \right) \rightarrow 0$ as $\lambda \rightarrow \infty$. Therefore, the condition (48) follows from the bound in (47).

(B16.18) It is sufficient to show that for $\gamma > 0$ and $\eta > 0$ arbitrarily small, there exists a δ small enough and a λ_0 large enough such that

$$P \left(\sup_{0 \leq t \leq T - \delta} \sup_{v, s \in [t, t + \delta]} \left| \tilde{Q}^\lambda(s) - \tilde{Q}^\lambda(v) \right| \geq \gamma \right) < \eta, \quad \lambda \geq \lambda_0. \quad (49)$$

Without loss of generality, assume $s < v$. We require an upper and a lower bound on the process $Q^\lambda(v) - Q^\lambda(s)$. For the upper bound, define a comparison system

$$Q_C^\lambda(t) \equiv A^\lambda(s + t) - A^\lambda(s) - \mu(\lambda) \int_s^t \mathbf{1}\{Q_C^\lambda(\zeta) > 0\} d\zeta.$$

By similar reasoning as in the previous paragraph,

$$Q^\lambda(v) - Q^\lambda(s) \leq Q_C^\lambda \left(\left\lfloor \frac{v - s}{l^\lambda} \right\rfloor l^\lambda \right) + A^\lambda(v) - A^\lambda \left(\left\lfloor \frac{v}{l^\lambda} \right\rfloor l^\lambda \right). \quad (50)$$

For the lower bound, since

$$\begin{aligned} & Q^\lambda(v) - Q^\lambda(s) \\ &= A^\lambda(v) - A^\lambda \left(\left\lfloor \frac{v}{l^\lambda} \right\rfloor l^\lambda \right) + A^\lambda \left(\left\lfloor \frac{v}{l^\lambda} \right\rfloor l^\lambda \right) - A^\lambda(s) + A^\lambda \left(\left\lceil \frac{s}{l^\lambda} \right\rceil l^\lambda \right) - A^\lambda \left(\left\lceil \frac{s}{l^\lambda} \right\rceil l^\lambda \right) \\ & \quad - S_I^\lambda(v) + S_I^\lambda(s) - S_O^\lambda(v) + S_O^\lambda(s) - N \left(\int_0^v \gamma Q_O^\lambda(\zeta) d\zeta \right) + N \left(\int_0^s \gamma Q_O^\lambda(\zeta) d\zeta \right), \end{aligned}$$

and at most n^λ customers are served every l^λ time units,

$$Q^\lambda(v) - Q^\lambda(s) \geq \sum_{i=\lceil s/l^\lambda \rceil}^{\lfloor v/l^\lambda \rfloor} (A^\lambda((i+1)l^\lambda) - A^\lambda(il^\lambda) - n^\lambda) - 2n^\lambda \quad (51)$$

$$- N \left(\int_0^v \gamma Q_O^\lambda(\zeta) d\zeta \right) + N \left(\int_0^s \gamma Q_O^\lambda(\zeta) d\zeta \right).$$

Noting that $\tilde{Q}_O^\lambda(t) \leq \tilde{Q}^\lambda(t)$ for all $t \geq 0$,

$$\frac{1}{\sqrt{\lambda}} (A^\lambda((i+1)l^\lambda) - A^\lambda(il^\lambda) - n^\lambda) = \tilde{A}^\lambda((i+1)l^\lambda) - \tilde{A}^\lambda(il^\lambda) - \lambda^{-2/3}\theta,$$

and

$$\frac{1}{\sqrt{\lambda}} \left(N \left(\int_0^v \gamma Q_O^\lambda(\zeta) d\zeta \right) - N \left(\int_0^s \gamma Q_O^\lambda(\zeta) d\zeta \right) \right)$$

$$= \tilde{N}^\lambda(\bar{\tau}^\lambda(v)) - \tilde{N}^\lambda(\bar{\tau}^\lambda(s)) - \int_s^v \gamma \tilde{Q}_O^\lambda(\zeta) d\zeta,$$

it follows from (50) and (51) that

$$\left| \tilde{Q}^\lambda(v) - \tilde{Q}^\lambda(s) \right| \leq \max(M_U^\lambda, M_L^\lambda), \quad (52)$$

where

$$M_U^\lambda \equiv \tilde{Q}_C^\lambda \left(\left\lfloor \frac{v-s}{l^\lambda} \right\rfloor l^\lambda \right) + \left| \tilde{A}^\lambda(v) - \tilde{A}^\lambda \left(\left\lfloor \frac{v}{l^\lambda} \right\rfloor l^\lambda \right) \right| + \sqrt{\lambda} \left(v - \left\lfloor \frac{v}{l^\lambda} \right\rfloor l^\lambda \right)$$

$$M_L^\lambda \equiv \left| \tilde{A}^\lambda \left(\left\lfloor \frac{v}{l^\lambda} \right\rfloor l^\lambda \right) - \tilde{A}^\lambda \left(\left\lceil \frac{s}{l^\lambda} \right\rceil l^\lambda \right) \right| + \left| \tilde{N}^\lambda(\bar{\tau}^\lambda(v)) \right| + \left| \tilde{N}^\lambda(\bar{\tau}^\lambda(s)) \right|$$

$$+ \gamma(v-s) \sup_{0 \leq t \leq T} \left| \tilde{Q}^\lambda(t) \right| + \lambda^{-2/3}\theta \left(\left\lfloor \frac{v}{l^\lambda} \right\rfloor l^\lambda - \left\lceil \frac{s}{l^\lambda} \right\rceil l^\lambda \right) + 2 \frac{n^\lambda}{\sqrt{\lambda}}.$$

The condition (49) follows because every term on the right-hand side of (52) becomes arbitrarily small with high probability as δ converges to 0, for $|v-s| < \delta$ and large enough λ . In particular, $\lfloor (v-s)/l^\lambda \rfloor l^\lambda \rightarrow v-s$ as $\lambda \rightarrow \infty$, and so, since \tilde{Q}_C^λ weakly converges to a continuous limit process (a reflected Brownian motion) with initial position 0, it follows that $\tilde{Q}_C^\lambda(\lfloor (v-s)/l^\lambda \rfloor l^\lambda)$ can be made arbitrarily small with high probability as δ becomes small. Furthermore, \tilde{A}^λ converges to a continuous limit process and so the terms

$$\left| \tilde{A}^\lambda(v) - \tilde{A}^\lambda \left(\left\lfloor \frac{v}{l^\lambda} \right\rfloor l^\lambda \right) \right| \quad \text{and} \quad \left| \tilde{A}^\lambda \left(\left\lfloor \frac{v}{l^\lambda} \right\rfloor l^\lambda \right) - \tilde{A}^\lambda \left(\left\lceil \frac{s}{l^\lambda} \right\rceil l^\lambda \right) \right|$$

become arbitrarily small with high probability as δ becomes small. The constant terms all converge to 0, and, because we have shown the convergence in (21) in Lemma 4, which implies (18) remains valid in this setting, $\tilde{N}^\lambda \circ \bar{\tau}^\lambda$ weakly converges to 0. Finally, because we have already shown condition (B16.17) is satisfied, the term $\gamma(v-s) \sup_{0 \leq t \leq T} |\tilde{Q}^\lambda(t)|$ becomes arbitrarily small with high probability as δ becomes small.

Proof of (19) and (20) (Lemma 3 equivalent)

As in the proof of Lemma 3, it is sufficient to show that as $\lambda \rightarrow \infty$

$$\tilde{P}_I^\lambda \Rightarrow \frac{\alpha w_O}{(1-\alpha)w_I + \alpha w_O} \tilde{Q} \text{ and } \tilde{P}_O^\lambda \Rightarrow \frac{(1-\alpha)w_I}{(1-\alpha)w_I + \alpha w_O} \tilde{Q}.$$

For the inline queue, note that the number of batches required to serve all customers in the inline queue exceeds $\lfloor Q_I^\lambda(t)/n^\lambda \rfloor$ and is less than $\lceil Q_I^\lambda(t)/n^\lambda \rceil$. Since each batch requires l^λ time units to process

$$l^\lambda \left\lfloor \frac{Q_I^\lambda(t)}{n^\lambda} \right\rfloor \leq P_I^\lambda(t) \leq l^\lambda \left\lceil \frac{Q_I^\lambda(t)}{n^\lambda} \right\rceil,$$

and so

$$0 \leq \sqrt{\lambda} P_I^\lambda(t) - \sqrt{\lambda} l^\lambda \left\lfloor \frac{Q_I^\lambda(t)}{n^\lambda} \right\rfloor \leq \sqrt{\lambda} l^\lambda.$$

Since $\sqrt{\lambda} l^\lambda \rightarrow 0$ as $\lambda \rightarrow \infty$ and by parts (i) and (ii) of this Proposition the weak convergence in Corollary 1 remains valid,

$$\sqrt{\lambda} l^\lambda \frac{Q_I^\lambda}{n^\lambda} = \frac{\lambda l^\lambda}{n^\lambda} \tilde{Q}_I^\lambda \Rightarrow \frac{\alpha w_O}{(1-\alpha)w_I + \alpha w_O} \tilde{Q}.$$

We conclude

$$\tilde{P}_I^\lambda \Rightarrow \frac{\alpha w_O}{(1-\alpha)w_I + \alpha w_O} \tilde{Q}$$

as $\lambda \rightarrow \infty$.

Since whenever the number of customers in the offline queue exceeds $(1-\alpha)n^\lambda$ at a discrete review time point, at least $(1-\alpha)n^\lambda$ customers are served,

$$\left(\frac{Q_O^\lambda(t)}{(1-\alpha)n^\lambda} + 1 \right) l^\lambda$$

exceeds the amount of time required for all customers in the offline queue that do not abandon to be served. Hence the number of customers in the offline queue that eventually do abandon

must be less than or equal to

$$\mathcal{A}^\lambda(t) \equiv N \left(\int_0^{t + \left(\frac{Q_O^\lambda(t)}{(1-\alpha)n^\lambda} + 1 \right) l^\lambda} \gamma Q_O^\lambda(s) ds \right) - N \left(\int_0^t \gamma Q_O^\lambda(s) ds \right).$$

Therefore,

$$l^\lambda \left[\frac{Q_O^\lambda(t) - \mathcal{A}^\lambda(t)}{n^\lambda} \right] \leq P_O^\lambda(t) \leq l^\lambda \left[\frac{Q_O^\lambda(t)}{n^\lambda} \right].$$

It follows from the observation that

$$\left(\frac{Q_O^\lambda(t)}{(1-\alpha)n^\lambda} + 1 \right) l^\lambda = \frac{Q_O^\lambda(t)}{(1-\alpha)\mu(\lambda)} + l^\lambda \rightarrow 0$$

as $\lambda \rightarrow \infty$ that

$$\frac{1}{\sqrt{\lambda}} \mathcal{A}^\lambda \Rightarrow 0$$

as $\lambda \rightarrow \infty$ by identical argument as that in the proof of Lemma 3. As in the preceding paragraph, we conclude

$$\tilde{P}_O^\lambda \Rightarrow \frac{(1-\alpha)w_I}{(1-\alpha)w_I + \alpha w_O} \tilde{Q}$$

as $\lambda \rightarrow \infty$.

□