Managing Service Systems with an Offline Waiting Option and Customer Abandonment

Vasiliki Kostami • Sriram Dasu • Amy R. Ward

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

February 6, 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. Next, we use the one-dimensional limit process to develop approximations for the steady-state distribution of the number of customers waiting inline and offline, the steady-state probability of abandonment from the offline queue, and to dynamically estimate inline and offline waits for each arriving customer. We conclude by considering a cost model, and optimize the upfront capacity allocation decision.

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 Disney 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 particular importance to investigate service models in which customers can choose between inline and offline queueing at the time of their arrival. Operating such a system requires that the server provides arriving customers with waiting time estimates 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 dynamically estimate wait times for customers as a function of observed queue-lengths. The difficulty inherent in making such estimates accurately is complicated by the presence of customers in the offline queue that may abandon. The wait time estimates we propose depend on an upfront static decision of how to allocate server capacity between the inline and 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 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 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 $\mu(1-\alpha)$. We later extend our analysis to include discrete review batch service that mimics the operation of a ride at an amusement park.

Our approach is to study the asymptotic behavior of the aforementioned system when demand is large (for example, hundreds of customers arrivals per hour) and service is fast (for example, the service time of one customer is measured in minutes). In this heavy-traffic asymptotic regime, the two dimensional process tracking the number of customers waiting inline and offline collapses to one dimension. This state-space collapse parallels the result in Theorem 1 in Section 5 in Reiman (1984) for a join-the-shorter queueing model in which no customers abandon. However, our one-dimensional limit process is a reflected Ornstein-Uhlenbeck process (which has state-dependent linear drift), whereas the one-dimensional limit process in Theorem 2 in Section 5 in Reiman (1984) is a reflected Brownian motion (which has constant drift). This is significant because both the steady-state and the transient behavior of the two processes are much different (Ward and Glynn 2003). For example, it is well-known that the steady-state distribution of a reflected Brownian motion is exponential but that of a reflected Ornstein-Uhlenbeck is truncated normal.

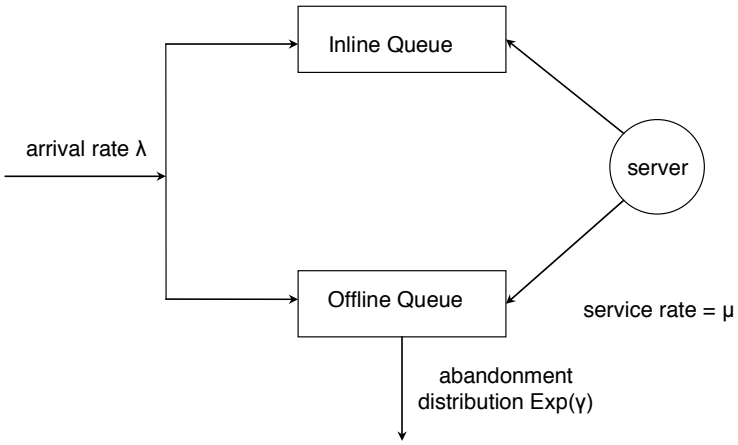

Figure 1: The model

We use the one-dimensional limiting reflected Ornstein-Uhlenbeck process to dynamically estimate inline and offline waits for each arriving customer. Additionally, it is straightforward to provide approximations for other performance measures of interest, such as the steady-state distribution of the number of customers in the inline and offline queues, and the steady-state probability of abandonment from the offline queue. We evaluate our proposed approximations using simulation, and find them to be very accurate when the probability a customer that chooses to wait offline abandons is under 10%.

Of course, as mentioned previously, the performance metrics depend on how capacity is allocated to each line. Allocation of capacity is therefore an important decision variable. Hence we further provide a cost model, and use the very analytically tractable approximating one-dimensional limiting diffusion to solve an optimization problem that determines how to allocate capacity between the inline and offline queue.

The remainder of the paper is organized as follows. We first review some relevant literature. In Section 2, we present our basic model formulation and heavy traffic asymptotic regime. We perform our asymptotic analysis in Section 3. We develop approximations for the discrete event system and test these approximations through simulation in Section 4. In Section 5, we extend our analysis to include discrete-review service in which customers are served in batches. Finally, in Section 6, we provide a cost model and optimize the capacity allocation decision.

Literature Review

The model we analyze can be viewed as a variant of a join-the-shorter queue model with real-time delay quotation. Hence the relevant literature falls primarily into two categories: papers that address delay (or leadtime) quotation in queueing models and papers that analyze join-the-shorter queue type models. In the first category, a common formulation (see, for example, Bookbinder and Noor 1985, Keskinocak et al 2001, Hopp and Sturgis 2001, Spearman and Zhang 1999, Wein 1991) is to minimize average leadtimes subject to some given service constraint (such as the percentage of customers not served within their quoted leadtime). The tension these papers model is between having long leadtime quotes that discourage customers and causing customer dissatisfaction by not serving customers within a promised leadtime. However, as discussed in Whitt (1999a) for a multi-server system without abandonments and in Whitt (1999b) for a multi-server system with abandonments in the form of both balking and reneging, the problem of accurately predicting queueing delays in service systems is of interest in its own right. Furthermore, having accurate delay prediction improves customer satisfaction. In contrast to Whitt (1999a)(1999b), in which delay prediction results often rely on Laplace transformation inversion methods, our focus is on providing a delay quotation policy for both the inline and offline queues, based on observed queue sizes, that is asymptotically compliant in the sense of Plambeck, Kumar, and Harrison (2001). That is, the delay quotes we provide to arriving customers coincide with the waiting times customers actually experience in our asymptotic regime. We expect from the work of Puhalskii (1994) that a delay quote for the offline queue that is formed by multiplying the number of customers in the offline queue that will eventually receive service with the average customer service requirement asymptotically coincides with actual waiting times. The difficulty is that our delay quotations must be based on the total number of customers in the offline queue, because we do not know which ones will abandon and which will not.

In the second category, one of the earlier works on the join-the-shorter queue model is that of Flatto and McLean (1977), who obtained an exact solution for the generating function of the stationary distribution in an exponential model having identical service rates for each queue. Adan et al (1991) extended this analysis to the asymmetric case. The earliest heavy traffic results on the join-the-shorter queue model are by Foschini and Salz (1978) for an exponential model. Results for a join-the-shorter queue model with general inter-arrival and service times can be found in Reiman (1984) (along with results on several other models that show state space collapse in a heavy traffic asymptotic regime). More recent work (McDonald 1996, Turner 2000) examines the large deviations limit. However, none of this literature allows for customer abandonments.

2. Model Formulation

We develop a sequence of generalized join-the-shorter queue queueing systems in which customer demand becomes large and service is rendered quickly, so that server utilization approaches unity. In Subsection 2.1, we present our basic system model. In Subsection 2.2, we describe our heavy traffic asymptotic regime.

2.1 Basic System Model

The service provider quotes every arriving customer an estimation of the waiting time in the inline and offline queues. We let $\mathcal{W}_I(t)$ represent the quote for the inline queue at time $t > 0$ and $\mathcal{W}_O(t)$ represent the quote for the offline queue. We assume customers are homogeneous in their waiting time costs, and let w_I and w_O be the waiting costs per time unit 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.

Waiting time quotes must be a function of observed queue-lengths. Suppose that the length of the inline queue at time $t \geq 0$ is $Q_I(t)$, and that the length of the offline queue is $Q_O(t)$. We assume that each arriving customer generates in expectation a workload of μ^{-1}, and that the server operates at a unit rate. Then, the expected total amount of processing required by all the customers in the inline and offline queues is $\mu^{-1}Q_I(t)$ and $\mu^{-1}Q_O(t)$ respectively. When customers are present in both queues, the inline queue is processed at rate α, and the offline queue is processed at rate $1-\alpha$. Otherwise, if the inline (offline) queue is empty, the offline (inline) queue is processed at rate 1. The service provider quotes a waiting time at time t that coincides with the expected processing time required by all customers in the queue adjusted by the service rate for each queue, assuming that the queues will be sharing processing capacity. Specifically, we assume that

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

In general, we expect that the waiting time quote $\mathcal{W}_O(t)$ will be too high. This is because not all customers present in the offline queue will receive service. Specifically, each customer joining the offline queue independently abandons after an exponential amount of time with mean γ^{-1}. However, we will show that in our heavy traffic asymptotic regime,

even though the presence of customer abandonment affects our proposed queue-length and waiting time process approximations, such an overestimation is small, because the probability any individual customer in the offline queue abandons becomes small. (See Theorem 3 for theoretical support of this statement, and our simulation results in Section 4 for numeric support.)

Let A, S_I, and S_O be independent renewal processes having rates λ, μ, and μ respectively. $A(t)$ denotes the cumulative number of arrivals to the system in $[0,t]$. $S_I(t)$ and $S_O(t)$ denote respectively the cumulative number of departures from the inline and offline queues after the server has devoted t units of time to the queue working at rate 1. Let N be an independent, standard Poisson process. The evolution equations for Q_I and Q_O are

$$\begin{aligned} Q_I(t) &\equiv \sum_{i=1}^{A(t)} \mathbf{1}\{w_I \mathcal{W}_I(t_i-) \leq w_O \mathcal{W}_O(t_i-)\} - S_I(T_I(t)) && (1) \\ Q_O(t) &\equiv \sum_{i=1}^{A(t)} \mathbf{1}\{w_I \mathcal{W}_I(t_i-) > w_O \mathcal{W}_O(t_i-)\} - N\left(\int_0^t \gamma Q_O(s) ds\right) \\ &\quad -S_O(T_O(t)), && (2) \end{aligned}$$

where

$$\begin{aligned} T_I(t) &\equiv \int_0^t \frac{\alpha \mathbf{1}\{Q_I(s)>0\}}{\alpha \mathbf{1}\{Q_I(s)>0\} + (1-\alpha)\mathbf{1}\{Q_O(s)>0\}} ds && (3) \\ T_O(t) &\equiv \int_0^t \frac{(1-\alpha) \mathbf{1}\{Q_O(s)>0\}}{\alpha \mathbf{1}\{Q_I(s)>0\} + (1-\alpha)\mathbf{1}\{Q_O(s)>0\}} ds. && (4) \end{aligned}$$

Define

$$Q \equiv Q_I + Q_O.$$

We assume the server must work whenever customers are present, and so

$$I(t) \equiv \int_0^t \mathbf{1}\{Q(s)=0\} ds \quad (5)$$

is the cumulative server idletime. Then,

$$\begin{aligned} T_I(t) + T_O(t) + I(t) &= t && (6) \\ \int_0^\infty Q(t) dI(t) &= 0. && (7) \end{aligned}$$

2.2 The Heavy Traffic Asymptotic Regime

We consider a sequence of systems, indexed by n, in which the arrival and service rates in the n^{th} system are of order n. 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 n^{th} system by n.

Let $\{u_i, i \geq 1\}$, $\{v_i^O, i \geq 1\}$ and $\{v_i^I, i \geq 1\}$ be three independent, i.i.d. sequences of non-negative, mean 1 random variables having finite variance. Further assume that $\{v_i^O, i \geq 1\}$ and $\{v_i^I, i \geq 1\}$ all have the same distribution. The cumulative number of arrivals is

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

so that the arrival rate in the n-th system is $n\lambda^n$. The server in the n-th system serves with rate $n\mu^n$ so that the cumulative number of customers served from the inline queue after the server has worked at rate 1 for t time units is

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

and from the offline queue is

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

Define the fluid scaled quantities

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

and the diffusion scaled quantities

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

Note that the queue-lengths are scaled by $n^{-1/2}$ and waiting time estimates are scaled by $\sqrt{n}$ because queue-lengths become large and waiting times become small in our limit regime. This is consistent with the limit regime in Reed and Ward (2007).

As n increases,

$$\lambda^n \to \mu \text{ and } \mu^n \to \mu,$$

where $\mu \in \Re$. (Note the slight abuse of notation because μ is no longer the service rate

introduced in Section 2.1. The service rate in the nth system is $n\mu^n$.) Furthermore, λ^n and μ^n become close at rate $\sqrt{n}$; specifically,

$$\sqrt{n}(\lambda^n - \mu^n) \to \theta, \tag{8}$$

as $n \to \infty$, where $\theta \in \Re$.

The functional strong law of large numbers establishes

$$\left(\bar{A}^n, \bar{S}_I^n, \bar{S}_O^n\right) \to (0, 0, 0) \text{ a.s., u.o.c.,} \tag{9}$$

as $n \to \infty$. Here, the notation "a.s." denotes "almost surely", and "u.o.c.", "uniformly on compact sets". Also, note that we let 0 represent both the 0 process as in (9) and the number 0. The meaning should be clear from the context.

It is useful to observe that the processes $\tilde{A}^n$, $\tilde{S}_I^n$, and $\tilde{S}_O^n$ can be approximated by Brownian motion. For this, we require the following technicalities. All random variables are defined on a common probability space $(\Omega, \mathcal{F}, P)$. For each positive integer d, let $D([0,\infty), \Re^d)$ be the space of right continuous functions with left limits (RCLL) in $\Re^d$ having time domain $[0,\infty)$. We endow $D([0,\infty), \Re^d)$ 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 $(\Omega, \mathcal{F}, P)$ into $(D([0,\infty), \Re^d), M^d)$ 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 $(D([0,\infty), \Re^d), M^d)$ converge weakly to the probability measure on $(D([0,\infty), \Re^d), M^d)$ induced by the stochastic process ξ. Note that we suppress d from the notation unless necessary.

Let B_u, B_{v_I} and B_{v_O} be independent, standard Brownian motions. The functional central limit theorem and the assumed independence of the interarrival and service time sequences establish

$$\left(\tilde{A}^n, \tilde{S}_I^n, \tilde{S}_O^n\right) \Rightarrow \left(\sqrt{\lambda \text{var}(u_1)} B_u, \sqrt{\mu \text{var}(v_1)} B_{v_I}, \sqrt{\mu \text{var}(v_1)} B_{v_O}\right). \tag{10}$$

In addition to the functional strong law of large numbers and the functional central limit theorem, we also reference the continuous mapping, random time change, and converging together theorems. A convenient reference for these theorems is Billingsley (1999) or Whitt (2002). We also often use the notation e to denote the identity process

$$e(t) = t \text{ for all } t \geq 0.$$

3. Asymptotic Analysis

Our first theorem establishes that the two-dimensional process tracking the number of customers waiting in the inline and offline queues collapses to one-dimension in heavy traffic.

Theorem 1 *For any $T > 0$, $\sup_{0\leq t\leq T} |\frac{w_I}{\alpha}\tilde{Q}_I^n(t) - \frac{w_O}{1-\alpha}\tilde{Q}_O^n(t)| \to 0$ in probability as $n \to \infty$.*

The next step is to identify the one-dimensional limit process. In preparation, let B be a standard Brownian Motion. Let

$$\sigma^2 = \lambda var(u_1) + \mu var(v_1^I).$$

Define Z as the strong solution to the stochastic equation

$$Z(t) = \theta t - \gamma \frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} \int_0^t Z(s)ds + \sigma B(t) + L(t) \geq 0,\ t \geq 0, \tag{11}$$

where L is non-decreasing, $L(0) = 0$ and $\int_0^\infty Z(t)dL(t) = 0$. The constant θ appearing in equation (11) corresponds to the assumed imbalance between the arrival and service rates in (8), and the term $\gamma(1-\alpha)w_I[\alpha w_O + (1-\alpha)w_I]^{-1}$ arises from customers abandonments from the offline queue.

The existence of a strong solution to (11) follows because the process Z can be represented in terms of the following regulator mapping.

Definition 1 *(The one-sided linearly generalized regulator mapping)*
Given κ a non-negative constant and $x \in D([0,\infty), \Re)$ having $x(0) \geq 0$, the one-sided linearly generalized regulator mapping

$$(\phi^\kappa, \psi^\kappa) : D([0,\infty), \Re) \mapsto D([0,\infty), [0,\infty) \times [0,\infty))$$

is defined by

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

where

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

(C2) l is nondecreasing, $l(0) = 0$, 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

$$(Z, L) = (\phi^\kappa, \psi^\kappa)\,(e + \sigma B). \tag{12}$$

Proposition 3 part (i) in Reed and Ward (2007) establishes the existence and uniqueness of the regulator mapping in Definition 1[1], and so the representation (12) guarantees that there is a unique strong solution to the stochastic equation in (11). Note that when $\kappa = 0$, the one-sided linearly generalized regulator mapping is exactly the conventional one-sided regulator mapping

$$\begin{aligned} \phi(x)(t) &\equiv x(t) + \psi(x)(t) \\ \psi(x)(t) &\equiv \sup_{s\in[0,t]} \max\{-x(s), 0\} \end{aligned}$$

introduced in Skorokhod (1961).

Our next theorem establishes that the process Z in (11) approximates the total number of customers in either the inline or offline queues.

Theorem 2 *As $n \to \infty$,*

$$\left(\tilde{Q}^n, \tilde{I}^n\right) \Rightarrow (Z, L).$$

Together, Theorems 1 and 2 imply a separate approximation for the number of customers in the inline queue, and for the number of customers in the offline queue. In particular, the following weak convergence holds

$$\tilde{Q}_I^n \Rightarrow \frac{\alpha w_O}{(1-\alpha)w_I + \alpha w_O} Z \text{ and } \tilde{Q}_O^n \Rightarrow \frac{(1-\alpha)w_I}{(1-\alpha)w_I + \alpha w_O} Z \tag{13}$$

as $n \to \infty$.

Our basic system model presented in Section 2.1 relies on the assumption that the amount of time a customer joining either the inline or offline queue (assuming he does not abandon) would wait to receive service at time t can be approximated from the queue-length processes. It is not obvious that the queue-length of the offline queue can be used to estimate waiting times because the number of customers that will abandon is not known. However, our next theorem shows that such an approximation is possible in our heavy traffic asymptotic regime.

Let W_I^n and W_O^n be 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

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

queue that will eventually receive service when processing occurs at rate 1. Let V_I^n and V_O^n be the virtual waiting time processes in the inline and offline queues respectively. The virtual waiting time processes multiplied by α and $1-\alpha$ respectively coincide with the workload processes when the server works continuously at rate α on the inline queue and rate $1-\alpha$ on the offline queue. Define $\tilde{W}_I^n = \sqrt{n}W_I^n$, $\tilde{W}_O^n = \sqrt{n}W_O^n$, $\tilde{V}_I^n = \sqrt{n}V_I^n$, $\tilde{V}_O^n = \sqrt{n}V_O^n$.

Theorem 3 *As $n \to \infty$,*

$$\tilde{W}_I^n \Rightarrow \frac{\alpha w_O}{(1-\alpha)w_I + \alpha w_O}\frac{Z}{\mu} \text{ and } \tilde{W}_O^n \Rightarrow \frac{(1-\alpha)w_I}{(1-\alpha)w_I + \alpha w_O}\frac{Z}{\mu}. \tag{14}$$

Furthermore, for any $T > 0$, as $n \to \infty$,

$$\sup_{0\le t\le T}\left|\tilde{V}_I^n(t) - \frac{\tilde{W}_I^n(t)}{\alpha}\right| \to 0 \text{ and } \sup_{0\le t\le T}\left|\tilde{V}_O^n(t) - \frac{\tilde{W}_O^n(t)}{1-\alpha}\right| \to 0, \tag{15}$$

in probability, which implies that, as $n \to \infty$,

$$\sup_{0\le t\le T}\left|\tilde{V}_I^n(t) - \tilde{\mathcal{W}}_I^n(t)\right| \to 0 \text{ and } \sup_{0\le t\le T}\left|\tilde{V}_O^n(t) - \tilde{\mathcal{W}}_O^n(t)\right| \to 0,$$

in probability.

Note that delay quotations are based on observed queue-lengths, which include abandoning customers, but that the workload and virtual waiting time processes do not include abandoning customers. Hence in our asymptotic regime it is possible to provide accurate delay quotations based on observed queue-lengths.

4. Approximations for the Original Model

In this section we develop approximations for steady-state performance measures for the original model, and test their accuracy through simulation (performed using the Extend simulation language). In Subsection 4.1, we use Theorems 1 through 3 to find closed-form expressions for expected steady-state queue-lengths, wait times, and the probability of abandonment from the offline queue and from the system. We show the results of our simulation study in Subsection 4.2.

4.1 Steady-State Approximations

Theorems 1 through 3 suggest that for any positive integer m

$$\begin{aligned} E\left[Q_I^n(\infty)^m\right] &\approx \sqrt{n}\frac{\alpha w_O}{(1-\alpha)w_I+\alpha w_O}E[Z(\infty)^m], \\ E\left[Q_O^n(\infty)^m\right] &\approx \sqrt{n}\frac{(1-\alpha)w_I}{(1-\alpha)w_I+\alpha w_O}E[Z(\infty)^m] \\ E\left[W_I^n(\infty)^m\right] &\approx \frac{1}{\sqrt{n}}\frac{\alpha w_O}{(1-\alpha)w_I+\alpha w_O}\frac{E[Z(\infty)^m]}{\mu} \\ E\left[W_O^n(\infty)^m\right] &\approx \frac{1}{\sqrt{n}}\frac{(1-\alpha)w_I}{(1-\alpha)w_I+\alpha w_O}\frac{E[Z(\infty)^m]}{\mu}, \end{aligned} \tag{16}$$

where $Q_I^n(\infty)$, $Q_O^n(\infty)$, $W_I^n(\infty)$, $W_O^n(\infty)$, and $Z(\infty)$ are random variables that have the steady-state distribution of the processes Q_I^n, Q_O^n, and Z respectively[2]. Next, 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

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

as in Section 3, the first and second moments of the steady-state distribution of the process Z are

$$\begin{aligned} E\left[Z(\infty)\right] &= \frac{\theta}{\kappa}+\frac{\sigma}{\sqrt{2\kappa}}\frac{\phi\left(\frac{-\theta}{\sigma}\sqrt{\frac{2}{\kappa}}\right)}{1-\Phi\left(\frac{-\theta}{\sigma}\sqrt{\frac{2}{\kappa}}\right)} \\ E\left[Z(\infty)^2\right] &= \left(\frac{\theta}{\kappa}\right)^2+\frac{\sigma^2}{2\kappa}+\frac{\theta\sigma}{\kappa\sqrt{2\kappa}}\frac{\phi\left(\frac{-\theta}{\sigma}\sqrt{\frac{2}{\kappa}}\right)}{1-\Phi\left(\frac{-\theta}{\sigma}\sqrt{\frac{2}{\kappa}}\right)}. \end{aligned}$$

Let p_O^n denote the steady-state probability that a customer who joins the offline queue abandons. To approximate p_O^n, first observe that the probability an infinitely patient customer arriving to the offline queue abandons when his wait for service is w is

$$1-\exp\left(-\gamma w\right)$$

Hence, because $W_O^n(t)/\mu(1-\alpha)$ is the approximate waiting time for service at the offline queue at time t, we expect that, for large t, assuming the process W_O^n is operating in steady

[2]Note that the proposed approximations assume limits can be taken either first as $n\to\infty$ and then as $t\to\infty$ or first as $t\to\infty$ and then $n\to\infty$, which has been proves rigorously for feedforward Jackson networks by Gamarnik and Zeevi (2006).

state,

$$\frac{\int_0^t \left(1-\exp\left(-\gamma\frac{W_O^n(s)}{\mu(1-\alpha)}\right)\right) dA^n(s)}{A^n(t)} = \frac{\int_0^t \left(1-\exp\left(-\gamma\frac{W_O^n(s)}{\mu(1-\alpha)}\right)\right) d(A^n(s)/\lambda^n n)}{A^n(t)/\lambda^n n}$$

is close to the steady-state abandonment probability for a customer joining the offline queue. Note that $A^n/n\lambda^n \to e$ u.o.c., a.s. as $n \to \infty$ and that for any $x \in \Re$,

$$\sqrt{n}\left(1-\exp\left(-\frac{x}{\sqrt{n}}\right)\right) \to x$$

$n \to \infty$. Then, multiplying by $\sqrt{n}$ (because the probability an individual customer abandons becomes small in our heavy traffic asymptotic regime) and using Theorem 3 suggests that

$$\sqrt{n}\frac{\int_0^t \left(1-\exp\left(-\gamma\frac{W_O^n(s)}{\mu(1-\alpha)}\right)\right) d(A^n(s)/\lambda^n n)}{A^n(t)/\lambda^n n} \to \frac{\gamma}{\mu^2(1-\alpha)}\frac{(1-\alpha)w_I}{(1-\alpha)w_I+\alpha w_O}\frac{\int_0^t Z(s)ds}{t},$$

as $n \to \infty$. (This argument can be made rigorous using the Skorohod representation theorem and Lemma 8.3 in Dai and Dai (1999), as shown in a similar argument in Subsection 5.1.2 in Reed and Ward (2007).) Finally, the strong law for regenerative processes implies that

$$\frac{\int_0^t Z(s)ds}{t} \to E\left[Z(\infty)\right]$$

as $t \to \infty$. We conclude that

$$p_O^n \approx \frac{1}{\sqrt{n}}\frac{\gamma}{\mu^2(1-\alpha)}\frac{(1-\alpha)w_I}{(1-\alpha)w_I+\alpha w_O}E\left[Z(\infty)\right]. \tag{17}$$

It is also possible to approximate the probability that an arbitrary customer arriving to the system abandons. In heavy-traffic, the inline and the offline queues are rarely empty, meaning that the inline queue receives α of the server's processing capacity and the offline queue receives $1-\alpha$. Consequently the fraction of arriving customers who join the offline must equal

$$\frac{1-\alpha}{(1-\alpha)+\alpha(1-p_O^n)}$$

and so

$$\frac{1}{(1-\alpha)+\alpha(1-p_O^n)}\frac{\gamma}{\sqrt{n}\mu^2}\frac{(1-\alpha)w_I}{(1-\alpha)w_I+\alpha w_O}E\left[Z(\infty)\right]$$

represents the probability that an arbitrary customer arriving to the system abandons.

4.2 Evaluation of the Approximations

Our first simulation study investigates the accuracy of our approximations in equations (16) and (17) as arrival and service rates become large. Specifically, Table 1 considers a balanced system ($\alpha = 0.5$) in which customers are indifferent between the two modes of service ($w_I = w_O = 1$). Then we expect the inline and offline queue-lengths to be equal, and so the reported % error for the first and second moment approximations of the number of customers in the inline and offline queue is the maximum of the errors associated with the two simulated values. The accuracy of the first and second moment approximations for the number of customers in the inline and offline queues is high (under 10%) when p_O^n is under 0.1, and the accuracy of the second moment approximation is high when p_O^n is under 0.05. Because the steady-state distribution of the limit process Z is a truncated normal, a good approximation of the first and second moments of the queue-length process suggests that their entire steady-state distribution is well-approximated.

	p_O^n		First Moment		Second Moment	
n	$\frac{\gamma}{\sqrt{n}}E[Z(\infty)]$	Error	$\frac{\sqrt{n}}{2}E[Z(\infty)]$	Max.Error	$\frac{\sqrt{n}}{2}E[Z(\infty)^2]$	Max.Error
10	0.252	5.99%	1.262	18.6%	2.5	54.6%
100	0.080	4.37%	3.989	8.3%	25	15.54%
1,000	0.025	0.01%	12.62	4.62%	250	7.22%
10,0000	0.008	4.09%	39.89	3.00%	2500	6.36%
100,0000	0.003	0.96%	126.16	0.96%	25,000	2.71%

Table 1: A comparison of the approximated offline queue abandonment probability, and the first and second moments of the number of customers in the inline and offline queues to a simulation having Poisson arrivals with rate n per time unit, deterministic service with mean $1/n$, and parameters $\alpha = 0.5$, $\gamma^{-1} = 1$, and $w_I = w_O = 1$.

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.

Recall that the waiting time approximations we provide to customers for the offline queue rely on assumption that estimated wait times for the offline queue can be approximated from the offline queue-length, and that Theorem 2 validates this assumption. One consequence of this assumption should be that the difference between the expected wait in the offline queue conditional on receiving service and the unconditional wait is very small when arrival and service rates are large. Table 2 verifies via simulation that this is indeed the case.

n	p_O^n	Approximated Wait $\frac{E[Z(\infty)]}{\sqrt{n}}$	Simulated Wait (%Error) Serviced Customers	All Customers
10	0.252	0.252	0.268 (5.97%)	0.297 (14.92%)
100	0.080	0.080	0.083 (4.32%)	0.086 (6.73%)
1,000	0.025	0.025	0.0252 (0.08%)	0.0254 (0.5%)
10,0000	0.008	0.008	0.00772 (3.39%)	0.00773 (3.21%)
100,0000	0.003	0.003	0.00255 (1.00%)	0.00255 (1.02%)

Table 2: A comparison of the unconditional and conditional (on receiving service) waiting times in the offline queue to a simulation having Poisson arrivals with rate n per time unit, deterministic service with mean $1/n$, and parameters $\alpha = 0.5$, $\gamma^{-1}= 1$, and $w_I = w_O = 1$.

	p_O^n		E[inline queue-length]		E[offline queue-length]	
α	Approximated	Error	Approximated	Error	Approximated	Error
0.0	0.0056	5.57%	0.00	N/A	56.42	5.52%
0.1	0.0059	3.61%	5.95	0.21%	53.52	4.69%
0.2	0.0063	5.33%	12.62	4.54%	50.46	5.89%
0.3	0.0067	2.87%	20.23	3.47%	47.20	2.57%
0.4	0.0073	1.68%	29.13	1.20%	43.70	1.29%
0.5	0.0080	4.29%	39.89	6.36%	39.89	5.24%
0.6	0.0089	4.98%	53.52	3.69%	35.68	4.66%
0.7	0.0103	1.57%	72.10	2.88%	30.90	1.88%
0.8	0.0126	0.82%	100.92	3.41%	25.23	1.72%
0.9	0.0178	9.73%	160.57	5.82%	17.84	8.58%

Table 3: A comparison for varying α of the approximated probability of abandonment and expected number of customers in inline and offline queues to a simulation having Poisson arrivals with rate 100 per time unit, deterministic service with mean 0.01, and parameters $\gamma^{-1} = 0.01$, and $w_I = w_O = 1$.

Tables 1 and 2 are suggestive of the effect γ has on the accuracy of our approximations. Specifically, the accuracy increases as the mean abandonment time becomes large compared to service times. It is the ratio of the two rather than the exact value that is important. Hence we do not include a study of how the approximation accuracy varies with γ.

Finally, we investigate the impact of capacity allocation on the accuracy of our proposed approximation. Specifically, we vary α between 0 and 1. Table 3 shows that there is no correlation between the value of α and the accuracy of our proposed approximations. Note that the value $\alpha = 1$ is not included because the system is unstable without abandonments.

When $\theta = 0$, it is straightforward to show that

$$\frac{d}{d\alpha}\frac{\alpha w_O}{(1-\alpha)w_I+\alpha w_O}E[Z(\infty)]$$
$$=\frac{\sigma}{\sqrt{\pi}\sqrt{\gamma}}\frac{w_I w_O}{((1-\alpha)w_I+\alpha w_O)^2}\left(\begin{array}{l}\left(\frac{(1-\alpha)w_I}{\alpha w_O+(1-\alpha)w_I}\right)^{1/2}\\ +\frac{1}{2}\frac{\alpha w_O}{(1-\alpha)w_I+\alpha w_O}\left(\frac{(1-\alpha)w_I}{\alpha w_O+(1-\alpha)w_I}\right)\end{array}\right)^{-3/2}$$
$$>0$$

and

$$\frac{d}{d\alpha}\frac{(1-\alpha)w_I}{(1-\alpha)w_I+\alpha w_O}E[Z(\infty)]$$
$$=-\frac{\sigma}{2\sqrt{\pi}\sqrt{\gamma}}\frac{w_I w_O}{((1-\alpha)w_I+\alpha w_O)^2}\left(\frac{(1-\alpha)w_I}{\alpha w_O+(1-\alpha)w_I}\right)^{1/2}$$
$$<0.$$

Hence we expect that as more capacity is devoted to the offline queue (i.e., as α increases), the number of customers present in the ofline queue also increases. In other words, customers will tolerate longer inline queue-lengths before choosing to go offline when more server effort is devoted to the inline queue. This is because the same number of customers at the inline queue results in a shorter wait time when the server devotes more effort to the inline queue. Table 3 verifies this intuition.

5. A Batch Processor with Discrete Service Start Times

In some applications settings, a fixed number of customers are served at set time intervals. For example, 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 only at certain time points. We extend our analysis in Subsection 5.1, and, in Subsection 5.2, test its validity via simulation using parameters from a ride that opened in Summer 2006 at Six Flags Magic Mountain, Tatsu.

5.1 Discrete Review Model Formulation

We modify our service process to be discrete-review. Otherwise, the evolution equations for the queue-length processes in (1) and (2) continue to hold. We again consider a sequence of systems, indexed by n, in which the arrival process in the nth system has rate $n\lambda^n$, and

is defined exactly as in Subsection 2.2. In a slight abuse of notation, we again use S_I^n and S_O^n to denote the cumulative departure processes 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^n and S_O^n refer to the processes defined below.

Let

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

and assume service occurs only at discrete review time points $l^n, 2l^n, 3l^n, \ldots$. No service occurs in between discrete review time points. Assume $n^{1/3}\mu^n$ customers can be processed in l^n units of time so that the processing rate is $n^{1/3}\mu^n/l^n = n\mu^n$ customers per unit time.[3]

The service process is defined recursively as follows. At time 0, no customers have been serviced so that

$$S_I^n(0) = S_O^n(0) = 0.$$

Next, observe that when the cumulative number of customers processed up to time $(i-1)l^n$ in the inline and offline queues is $S_I^n((i-1)l^n)$ and $S_O^n((i-1)l^n)$ respectively, then because no customers are processed in between discrete review time points,

$$\begin{aligned}
Q_I^n(il^n-) &= Q_I^n((i-1)l^n) + \sum_{i=A^n((i-1)l^n)+1}^{A^n(il^n)} \mathbf{1}\left\{\frac{w_I}{\mu\alpha}Q_I^n(t_i^n-) \le \frac{w_O}{\mu(1-\alpha)}Q_O^n(t_i^n-)\right\} \\
Q_O^n(il^n-) &= Q_O^n((i-1)l^n) + \sum_{i=A^n((i-1)l^n)+1}^{A^n(il^n)} \mathbf{1}\left\{\frac{w_I}{\mu\alpha}Q_I^n(t_i^n-) > \frac{w_O}{\mu(1-\alpha)}Q_O^n(t_i^n-)\right\} \\
&\quad -N\left(\int_0^{il^n} \gamma Q_O^n(s)ds\right) + N\left(\int_0^{(i-1)l^n} \gamma Q_O^n(s)ds\right).
\end{aligned}$$

Then,

$$S_I^n(il^n) = \begin{cases} S_I^n((i-1)l^n) + \lfloor \alpha n^{1/3}\mu^n \rfloor & Q_I^n(il^n-) \ge \lfloor \alpha n^{1/3}\mu^n \rfloor \\ \quad + \left(\begin{array}{l} \left[\lceil (1-\alpha)n^{1/3}\mu^n \rceil - Q_O(il^n-)\right]^+ \\ \wedge \left(Q_I^n(il^n-) - \lfloor \alpha n^{1/3}\mu^n \rfloor\right) \end{array} \right) & \\ S_I^n((i-1)l^n) + Q_I^n(il^n-) & Q_I^n(il^n-) < \lfloor \alpha n^{1/3}\mu^n \rfloor \end{cases} \tag{18}$$

[3]The choice of $n^{-2/3}$ as a discrete review time period is somewhat arbitrary. The analysis in this section continues to hold for $l^n = n^{-\beta}$ for any $1/2 < \beta < 1$ when $n^{1-\beta}\mu^n$ is the number of customers processed in l^n units of time.

and

$$S_O^n(il^n) = \begin{cases} S_O^n((i-1)l^n) + \lceil (1-\alpha)n^{1/3}\mu^n \rceil & Q_O^n(il^n-) \geq \lceil (1-\alpha)n^{1/3}\mu^n \rceil \\ \quad + \left(\begin{array}{l} \left[\lfloor \alpha n^{1/3}\mu^n \rfloor - Q_I^n(il^n-) \right]^+ \\ \wedge \left(Q_O^n(il^n-) - \lceil (1-\alpha)n^{1/3}\mu^n \rceil \right) \end{array} \right) & \\ S_O^n((i-1)l^n) + Q_O^n(il^n-) & Q_O^n(il^n-) < \lceil (1-\alpha)n^{1/3}\mu^n \rceil \end{cases} \quad (19)$$

We expect that the discrete review system behaves similarly to the continuous time system. Our next proposition shows that Theorems 1 through 3 remain valid for the discrete review system. Its proof can be found in the appendix.

Proposition 1 *The results of Theorems 1 through 3 continue to hold when the service process is defined as in (18)-(19). The process Z appearing in Theorems 2 and 3 again solves the stochastic equation (11) but has infinitesimal variance $\sigma^2 = \lambda var(\mu_1)$.*

5.2 Application to the Tatsu Ride

Tatsu is a roller coaster ride at Magic Mountain Park. Each train in this ride has a capacity of 32 passengers, and approximately 1600 customers can take this ride in an hour (Tatsu 2007). Roughly once every 72 seconds a new train departs. We use our proposed approximations to predict queue-lengths when the Tatsu ride operates with both an inline and an offline queue.

	E[Queue Size]	Simulated $(E[Q_I^n], E[Q_O^n])$ / Max. Error	
$n\lambda^n$	$\sqrt{n}E[Z(\infty)]$	W/out Batching	With Batching
1600	31.92	(31.51, 31.01) / 2.93%	(36.53, 36.03) / 12.64%
1610	51.50	(52.44, 51.93) / 1.78%	(55.55, 55.04) / 7.29%
1620	82.21	(82.95, 82.43) / 0.89%	(87.01, 86.50) / 5.52%
1630	120.18	(126.53, 126.00) / 5.02%	(127.69, 127.17) / 5.88%
1640	160.01	(154.23, 153.70) / 4.10%	(158.82, 159.29) / 1.08%
1650	200.00	(200.34, 199.80) / 0.16%	(201.54, 201.01) / 0.76%

Table 4: A comparison of the expected inline and offline queue sizes to a simulation having Poisson arrivals at a rate $n\lambda^n$, service capacity $n\mu^n = 1600$ customers per hour, $\gamma^{-1} = 4$ hours, $\alpha = 0.5$, and $w_I = w_O = 1$ for (i) a system with continuous service, and (ii) a system with discrete-review service.

We parameterize the Tatsu ride in time units of hours as follows. Let $\mu^n = 1$ for all n, and $n = 1600$ so that the service rate is 1600 passengers per hour. The Tatsu ride was very popular the first summer it opened, and generally had long queues. Hence Table 4 investigates the performance of our approximations when the arrival rate to the ride $n\lambda^n$ exceeds the service rate. For the remaining parameters, we assume that the mean time a

customer in the offline queue spends in the park before deciding not to ride Tatsu is $\gamma^{-1} = 4$ hours, or approximately half the opening time of the park in a day. Furthermore, when there are enough customers in the two queues, exactly 16 passengers are taken from the inline queue and 16 from the offline queue so that $\alpha = 0.5$. Finally, customers are indifferent between the two modes of service $w_I = w_O = 1$.

Table 4 evidences the validity of our approximation for a discrete-review system. In particular, as established in Proposition 1, the difference between continuous and discrete-review service is small.

6. Revenue Optimization

For a designated split α of server effort between the inline and offline queues, we have developed approximations for the queue-length and waiting time processes when arrival and service rates are large, and performed simulation studies to verify their accuracy. We are now in a position to address the question: what is the optimum α? We conclude our paper by presenting an example cost model, and showing how to find the α that minimizes average cost.

There is a cost c associated with any customer that abandons. The cost could be a refund for a service not rendered or, as in the amusement park setting, could represent an expected future revenue loss due to the customer being less likely to return at a later date and pay for more service. Also, there is a holding cost $h_I \in \Re$ per customer in the inline queue, and a holding cost $h_O \in \Re$ per customer in the offline queue. Note that we allow h_I and h_O to be negative to allow for the case that customers in queue can generate revenue. For example, in an amusement park setting, customers in the offline queue may purchase food and spend money on entertainment, so that the holding cost h_O is actually a revenue generated per customer while wandering around the park. Then, the total cost after t time units in the system having arrival and service rates of order n is

$$\mathcal{C}^n(t) \equiv cN\left(\int_0^t \gamma Q_O^n(s)ds\right) + \int_0^t h_I Q_I^n(s)ds + \int_0^t h_O Q_O^n(s)ds.$$

Our objective is to minimize infinite horizon average cost

$$\min_{\alpha\in[0,1]} \lim_{t\to\infty} \frac{1}{t}\mathcal{C}^n(t).$$

(Note that we expect the limit to exist for any $\alpha \in [0,1]$ because we expect that the system is positive recurrent.) The problem as stated is intractable. However, when n is large, we

can utilize the approximations for the queue-length processes suggested by Theorem 2. To do this, because queue-lengths in the nth system are of order $\sqrt{n}$, $\mathcal{C}^n(t)$ is also of order $\sqrt{n}$ for any finite t. Hence we must divide $\mathcal{C}^n$ by $\sqrt{n}$.

An ideal alternative (tractable) objective is

$$\min_{\alpha\in[0,1]} \lim_{n\to\infty} \lim_{t\to\infty} \frac{1}{t}\frac{\mathcal{C}^n(t)}{\sqrt{n}}.$$

However, this alternative proves difficult because in order to use the approximating process Z, the limit must first be taken as $n \to \infty$ and second be taken as $t \to \infty$. Hence we utilize the alternative objective

$$\min_{\alpha\in[0,1]} \lim_{t\to\infty} \lim_{n\to\infty} \frac{1}{t}\frac{\mathcal{C}^n(t)}{\sqrt{n}}. \tag{20}$$

Observe that

$$\begin{aligned}\frac{\mathcal{C}^n(t)}{\sqrt{n}} &= c\left[\tilde{N}^n\left(\overline{\tau}^n(t)\right) + \gamma\int_0^t \tilde{Q}_O^n(s)ds\right] \\ &\quad + \int_0^t h_I\tilde{Q}_I^n(s)ds + \int_0^t h_O\tilde{Q}_O^n(s)ds.\end{aligned}$$

Define

$$\mathcal{C}(t) = \left[(c\gamma + h_O)\frac{(1-\alpha)w_I}{(1-\alpha)w_I + \alpha w_O} + h_I\frac{\alpha w_O}{(1-\alpha)w_I + \alpha w_O}\right]\int_0^t Z(s)ds.$$

By the same argument directly following (26) in the proof of Theorem 1, $\tilde{N}^n \circ \overline{\tau}^n \Rightarrow 0$, as $n \to \infty$. Hence, Theorems 1 and 2 (specifically, the weak convergence in (13)) and the continuous mapping theorem show

$$\frac{\mathcal{C}^n}{\sqrt{n}} \Rightarrow \mathcal{C},$$

as $n \to \infty$. As in Subsection 4.1

$$\frac{\int_0^t Z(s)ds}{t} \to E[Z(\infty)],$$

where $Z(\infty)$ has the steady-state distribution of the process Z given in (17). We conclude that the objective in (20) is equivalently expressed as

$$\min_{\alpha\in[0,1]} \left(\frac{(c\gamma + h_O)(1-\alpha)w_I}{(1-\alpha)w_I + \alpha w_O} + \frac{h_I\ \alpha w_O}{(1-\alpha)w_I + \alpha w_O}\right)\left(\frac{\theta}{\kappa} + \frac{\sigma}{\sqrt{2\kappa}}\frac{\phi\left(\frac{-\theta}{\sigma}\sqrt{\frac{2}{\kappa}}\right)}{1-\Phi\left(\frac{-\theta}{\sigma}\sqrt{\frac{2}{\kappa}}\right)}\right). \tag{21}$$

The optimization problem in (21) minimizes a continuous function over a bounded region, and so is solvable numerically.

For intuition, we solve (21) in the case that customer waiting costs are identical for each queue $w_I = w_O = 1$, and there is exact balance between the arrival and service rates so that $\theta = 0$. To relate back to the amusement park setting, we assume $h_I \geq 0, h_O < 0$, and $c\gamma + h_O > 0$. In this simplified setting, (21) becomes

$$\min_{\alpha \in [0,1]} f(\alpha), \tag{22}$$

where

$$f(\alpha) = \frac{\sigma}{\sqrt{\pi}\sqrt{\gamma}} \frac{1}{\sqrt{(1-\alpha)}} \left(c\gamma + h_O + \alpha(h_I - c\gamma - h_O)\right)$$

is a non-negative function in the interval $[0, 1]$. The function f has first derivative

$$f'(\alpha) = \frac{\sigma}{\sqrt{\pi}\sqrt{\gamma}}(1-\alpha)^{-3/2} \left(h_I - \frac{1}{2}(c\gamma + h_O) - \frac{1}{2}\alpha(h_I - c\gamma - h_O)\right),$$

and second derivative

$$f''(\alpha) = \frac{\sigma}{\sqrt{\pi}\sqrt{\gamma}}(1-\alpha)^{-5/2} \left(h_I - \frac{1}{4}(c\gamma + h_O) - \frac{1}{4}\alpha(h_I - c\gamma - h_O)\right).$$

To guarantee a solution $\alpha^\star \in (0, 1)$, we require the condition

$$0 < h_I < \frac{1}{2}\left(c\gamma + h_O\right). \tag{23}$$

When $h_I = 0$, $f(1) = 0$, the minimum achievable cost, and so $\alpha^\star = 1$. In other words, it is optimum to maintain only an inline queue. In the case that $h_I \geq 2^{-1}(c\gamma + h_O)$, it follows that $f'(\alpha) \geq 0$ for all $\alpha \in [0, 1]$. Then, the minimum achievable cost occurs at $\alpha^\star = 0$, and so having only an offline queue is optimum. Otherwise, when condition (23) is satisfied, solving $f'(\alpha) = 0$ shows

$$\alpha^\star = \frac{2\left(\frac{1}{2}(c\gamma + h_O) - h_I\right)}{c\gamma + h_O - h_I}.$$

Since $f'(0) < 0$ when $h_I < 2^{-1}(c\gamma + h_O)$, it follows that $f'(0) > f(\alpha^\star)$. Furthermore, $f(\alpha) \to \infty$ as $\alpha \uparrow 1$, and so $f(\alpha^\star) \leq f(\alpha)$ for all $\alpha \in [0, 1]$.

Finally, it is interesting to compare the solution to our optimization problem in (22) to the solution for the case that there is no abandonment. Then, 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), Theorems 1 and 2 hold except that

the process Z is a reflected Brownian motion with drift θ and variance σ^2. When $\theta < 0$, the steady-state mean of Z is σ^2/θ. (See, for example, equation (12) in Section 5.6 in Harrison (1985).) Hence, the objective (21) becomes

$$\min_{\alpha\in[0,1]} \frac{\sigma^2}{2|\theta|}\left(h_O(1-\alpha)+h_I\alpha\right)$$

when $w_I = w_O = 1$ and $\theta = 0$. The solution is "bang-bang": when $h_I < h_O$, the minimum occurs at $\alpha = 1$, and at $\alpha = 0$ when $h_I > h_O$[4]. We conclude that it is the presence of abandonments that causes the system manager to want to maintain both an inline queue and a offline queue.

Appendix

The proofs of Theorems 1-3 require the following two Lemmas, whose proofs we defer to the end of the appendix.

Lemma 1 *Let $W^n = W_I^n + W_O^n$. As $n \to \infty$,*

$$\left(\bar{Q}^n, \bar{W}^n, \bar{\tau}^n, \bar{T}_I^n + \bar{T}_O^n, \bar{I}^n\right) \to (0,0,0,e,0)\,,\ a.s.,\ u.o.c..$$

Lemma 2 *For any $T > 0$ and $\epsilon > 0$, there exists B and n_0 such that*

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

for all $n \ge n_0$.

Proof of Theorem 1

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 $\epsilon > 0$,

$$P\left(\sup_{0\le t\le T}\left|\frac{w_I}{\alpha}\tilde{Q}_I^n(t) - \frac{w_O}{1-\alpha}\tilde{Q}_O^n(t)\right| > \epsilon\right) \to 0 \text{ as } n \to \infty. \tag{24}$$

[4]The case that $h_I = h_O$ is degenerate in the sense that the cost function no longer depends on α, and so any $\alpha \in [0,1]$ achieves minimum cost.

Fix $\epsilon > 0$ and let

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

It will also be useful to define the processes

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

An upper bound for the left-hand-side of (24)

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

$$\begin{aligned}
&\left|\frac{w_I}{\alpha}\tilde{Q}_I^n(t) - \frac{w_O}{1-\alpha}\tilde{Q}_O^n(t)\right| \\
&= \frac{w_I}{\alpha}\tilde{Q}_I^n\left(\xi_n^*-\right) - \frac{w_O}{1-\alpha}\tilde{Q}_O^n\left(\xi_n^*-\right) - \frac{w_I}{\alpha}\frac{1}{\sqrt{n}}\left\{S_I^n(T_I^n(t)) - S_I^n(T_I^n(\xi_n^*-))\right\} \\
&\quad + \frac{w_O}{1-\alpha}\frac{1}{\sqrt{n}}\left\{S_O^n(T_O^n(t)) - S_O^n(T_O^n(\xi_n^*-))\right\} + \frac{w_O}{1-\alpha}\frac{1}{\sqrt{n}}N\left(\int_{\xi_n^*-}^t \gamma Q_O^n(s)ds\right) \\
&\quad - \frac{w_O}{1-\alpha}\left\{\tilde{A}^n(t) - \tilde{A}^n(\xi_n^*-) + \sqrt{n}\lambda^n(t-\xi_n^*)\right\}. \qquad (25)
\end{aligned}$$

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

$$T_I^n(t) - T_I^n(\xi_n^*-) \geq \alpha(t-\xi_n^*).$$

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

$$T_O^n(t) - T_O^n(\xi_n^*-) \le (1-\alpha)(t-\xi_n^*).$$

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

$$\begin{aligned} & S_I^n(T_I^n(t)) - S_I^n(T_I^n(\xi_n^*-)) \\ &\ge S_I^n(T_I^n(\xi_n^*-) + \alpha(t-\xi_n^*)) - S_I^n(T_I^n(\xi_n^*-)) \\ &= \sqrt{n}\left[\tilde{S}_I^n\left(T_I^n(\xi_n^*-) + \alpha(t-\xi_n^*)\right) - \tilde{S}_I^n\left(T_I^n(\xi_n^*-)\right) + \alpha\sqrt{n}(t-\xi_n^*)\right], \end{aligned}$$

and

$$\begin{aligned} & S_O^n(T_O^n(t)) - S_O^n(T_O^n(\xi_n^*-)) \\ &\le S_O^n(T_O^n(\xi_n^*-) + (1-\alpha)(t-\xi_n^*)) - S_O^n(T_O^n(\xi_n^*-)) \\ &= \sqrt{n}\left[\tilde{S}_O^n\left(T_O^n(\xi_n^*-) + (1-\alpha)(t-\xi_n^*)\right) - \tilde{S}_O^n\left(T_O^n(\xi_n^*-)\right) + (1-\alpha)\sqrt{n}(t-\xi_n^*)\right]. \end{aligned}$$

The definition of ξ_n^* and substitution of the above upper bounds into (25) establish

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

When $w_I\tilde{Q}_I^n(\xi_n^*)/(\mu\alpha) \le w_O\tilde{Q}_O^n(\xi_n^*)/(\mu(1-\alpha))$, a similar argument shows

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

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

$$\begin{aligned} & \left|\frac{w_I}{\alpha}\tilde{Q}_I^n(t) - \frac{w_O}{1-\alpha}\tilde{Q}_O^n(t)\right| \\ &\le \frac{\epsilon}{2} + \max\left\{\tilde{U}_1^n(t, \xi_n^*-, T_I^n(\xi_n^*-), T_O^n(\xi_n^*-)), \tilde{U}_2^n(t, \xi_n^*-, T_I^n(\xi_n^*-), T_O^n(\xi_n^*-)\right\} \\ &\quad + \frac{w_O}{1-\alpha}\frac{1}{\sqrt{n}}N\left(\int_0^T \gamma Q_O^n(s)ds\right). \end{aligned}$$

Therefore, the left-hand side of (24) can be bounded as follows

$$P\left(\sup_{0\le t\le T}\left|\frac{w_I}{\alpha}\tilde{Q}_I^n(t)-\frac{w_O}{1-\alpha}\tilde{Q}_O^n(t)\right|>\epsilon\right) \tag{26}$$
$$\le \quad P\left(\begin{array}{l}\sup_{0\le s\le t\le T}\sup_{0\le u,v\le s}\max\left\{\tilde{U}_1^n(t,s,u,v),\tilde{U}_2^n(t,s,u,v)\right\}\\ +\frac{w_O}{1-\alpha}\frac{1}{\sqrt{n}}\Pi^n\left(\int_0^T\gamma Q_O^n(s)ds\right)\end{array}>\frac{\epsilon}{2}\right).$$

Convergence of the right-hand-side of (26) to zero

Let η be arbitrarily small. Observe that

$$\frac{1}{\sqrt{n}}N\left(\int_0^T\gamma Q_O^n(s)ds\right)=\tilde{N}^n\left(\bar{\tau}^n(T)\right)+\gamma\int_0^T\tilde{Q}_O^n(s)ds$$

From Lemma 1,we know that $\bar{\tau}^n\to 0$ as $n\to\infty$ a.s., u.o.c. The functional central limit theorem establishes that $\tilde{N}^n$ weakly converges to a Brownian Motion as $n\to\infty$. Since τ^n is a non-decreasing process, the random time change theorem implies that $\tilde{N}^n\circ\bar{\tau}^n$ weakly converges to the zero process. Therefore, $\tilde{N}^n\left(\overline{\tau}^n(T)\right)\Rightarrow 0$ as $n\to\infty$. Since weak convergence to a constant is equivalent to convergence in probability and $\int_0^T\tilde{Q}_O^n(y)dy$ is stochastically bounded due to Lemma 2, there exists M and n_0 large enough so that

$$P\left(\frac{w_O}{1-\alpha}\frac{1}{\sqrt{n}}N\left(\int_0^T\gamma Q_O^n(s)ds\right)>M\right)<\frac{\eta}{2}$$

for all $n\ge n_0$.

The processes $\tilde{A}^n$, $\tilde{S}_I^n$, and $\tilde{S}_O^n$ all weakly converge to Brownian motions by the functional central limit theorem. The heavy traffic assumption (8) implies that for any $t>s$, as $n\to\infty$,

$$\begin{array}{rcl}\left(\frac{w_O}{1-\alpha}\left(\mu^n-\lambda^n\right)-\mu^n\left(w_I+\frac{\alpha}{1-\alpha}w_O\right)\right)\sqrt{n}(t-s) & \to & -\infty\\ \left(\frac{w_I}{\alpha}\left(\mu^n-\lambda^n\right)-\mu^n\left(\frac{1-\alpha}{\alpha}w_I+w_O\right)\right)\sqrt{n}(t-s) & \to & -\infty.\end{array}$$

Therefore, an argument analogous to the proof of Theorem 3.2 in Reiman (1984) shows that there exists m_0 such that for all $n>m_0$

$$P\left(\sup_{0\le s\le t\le T}\sup_{0\le u,v\le s}\max\left\{\tilde{U}_1^n(t,s,u,v),\tilde{U}_2^n(t,s,u,v)\right\}+M>\frac{\epsilon}{2}\right)<\eta.$$

We conclude that for all $n > n_0 \vee m_0$

$$P\left(\sup_{0\le s\le t\le T} \sup_{0\le u,v\le s} \max\left\{\tilde{U}_1^n(t,s,u,v), \tilde{U}_2^n(t,s,u,v)\right\} + \frac{w_O}{1-\alpha}\frac{1}{\sqrt{n}} N\left(\int_0^T \gamma Q_O^n(s)ds\right) > \frac{\epsilon}{2}\right)$$

$$\begin{aligned}
&\le P\left(\sup_{0\le s\le t\le T} \sup_{0\le u,v\le s} \max\left\{\tilde{U}_1^n(t,s,u,v), \tilde{U}_2^n(t,s,u,v)\right\} + M > \frac{\epsilon}{2}\right) \\
&\quad + P\left(\frac{w_O}{1-\alpha}\frac{1}{\sqrt{n}} N\left(\int_0^T \gamma Q_O^n(s)ds\right) > \frac{\epsilon}{2}\right) \\
&< \frac{\eta}{2} + \frac{\eta}{2} = \eta.
\end{aligned}$$

Proof of Theorem 2

Define

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

Then, for all $t \ge 0$,

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

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

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

it follows that

$$\left(\tilde{Q}^n, \mu^n\tilde{I}^n\right) \equiv (\phi^\kappa, \psi^\kappa)\left(\tilde{X}^n + \tilde{\epsilon}^n\right). \tag{27}$$

Let B be a standard Brownian motion. Suppose we can show

$$\tilde{X}^n \Rightarrow \sigma B + \theta e,$$

as $n \to \infty$. By the continuous mapping theorem and Theorem 1,

$$\tilde{\epsilon}^n \Rightarrow 0,$$

as $n \to \infty$. Proposition 4 part (iii) in Ward and Kumar (2007) establishes that the mapping

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

$$(\phi^\kappa, \psi^\kappa)\left(\tilde{X}^n + \tilde{\epsilon}^n\right) \Rightarrow (\phi^\kappa, \psi^\kappa)\left(\sigma B + \theta e\right),$$

as $n \to \infty$. The representation (Z, L) in terms of the one-sided linearly generalized regulator mapping in (27) shows $(Z, L) = (\phi^\kappa, \psi^\kappa)(\sigma B + \theta e)$, and so

$$\left(\tilde{Q}^n, \mu^n \tilde{I}^n\right) \Rightarrow (Z, L)$$

as $n \to \infty$.

The sequence $\{(T_O^n, T_I^n)\}$ is tight in D because $|T_I^n(t) - T_I^n(s)| \leq |t - s|$ and $|T_O^n(t) - T_O^n(s)| \leq |t - s|$. Consider any subsequence $\{n_k\}$ on which

$$(T_O^{n_k}, T_I^{n_k}) \Rightarrow (T_O, T_I)$$

as $n_K \to \infty$. 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 $\{n_k\}$, by the functional central limit theorem, continuous mapping theorem, and the heavy traffic assumption (8)

$$\begin{aligned} &\tilde{A}^{n_k}(t) - \tilde{S}_I^{n_k}\left(T_I^{n_k}(t)\right) - \tilde{S}_O^{n_k}\left(T_O^{n_k}(t)\right) + \sqrt{n_k} t\left(\lambda^{n_k} - \mu^{n_k}\right) \\ &\quad \Rightarrow \quad \sqrt{\mathrm{var}(u_1)} B_1 - \sqrt{\mathrm{var}(v_1^I)} B_2 \circ T_I - \sqrt{\mathrm{var}(v_1^O)} B_3 \circ T_O + \theta e, \end{aligned}$$

as $n_k \to \infty$. By the same argument directly following (26) in the proof of Theorem 1,

$$\tilde{N}^{n_k} \circ \overline{\tau}^{n_k}(t) \Rightarrow 0,$$

as $n_k \to \infty$. Therefore,

$$\tilde{X}^{n_k} \Rightarrow \sqrt{\mathrm{var}(u_1)} B_1 - \sqrt{\mathrm{var}(v_1^I)} B_2 \circ T_I - \sqrt{\mathrm{var}(v_1^O)} B_3 \circ T_O + \theta e,$$

as $n_k \to \infty$. Since $T_I + T_O = e$ and $\mathrm{var}(v_1^I) = \mathrm{var}(v_1^O)$ by assumption, it follows that $\mathrm{var}(u_1) B_1 - \mathrm{var}(v_1^I) B_2 \circ T_I - \mathrm{var}(v_1^O) B_3 \circ T_O$ has the same distribution as σB. Since the

subsequence $\{n_k\}$ was arbitrary, we conclude

$$\tilde{X}^n \Rightarrow \sigma B + \theta e,$$

as $n \to \infty$.

□

Proof of Theorem 3

Proof of (14)

We establish

$$\tilde{W}_O^n \Rightarrow \frac{(1-\alpha)w_I}{(1-\alpha)w_I + \alpha w_O} \frac{Z}{\mu} \tag{28}$$

as $n \to \infty$. Showing

$$\tilde{W}_I^n \Rightarrow \frac{\alpha w_O}{(1-\alpha)w_I + \alpha w_O} \frac{Z}{\mu}$$

as $n \to \infty$ follows an argument similar to Theorem 5.3 in Reiman (1984), and so is omitted. Since the offline service queue receives at least $(1-\alpha)$ proportion of the server's efforts when the queue is non-empty, $(1-\alpha)^{-1} W_O^n(t)$ exceeds 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

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

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

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

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

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

We conclude

$$0 \leq \sqrt{n}W_O^n(t) - \sqrt{n}L_O^n(t) \leq \sqrt{n}U_O^n(t) - \sqrt{n}L_O^n(t). \tag{29}$$

Define

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

Observe that

$$\begin{aligned} &\sqrt{n}U_O^n(t) - \sqrt{n}L_O^n(t) \\ &= \frac{1}{\mu^n}\frac{1}{\sqrt{n}} v_{S_O^n(T_O^n(t))+1} + \frac{1}{\mu^n}\frac{1}{\sqrt{n}}\mathcal{A}^n(t) \\ &\quad + \frac{1}{\mu^n}\left(\tilde{V}_O^n\left(\frac{S_O^n(T_O^n(t))}{n} + \frac{Q_O^n(t)}{n}\right) - \tilde{V}_O^n\left(\frac{S_O^n(T_O^n(t))}{n} + \frac{Q_O^n(t)}{n} - \frac{\mathcal{A}^n(t)}{n}\right)\right) \end{aligned} \tag{30}$$

and

$$\begin{aligned} &\sqrt{n}L_O^n(t) \\ &= \frac{1}{\mu^n}\tilde{Q}_O^n(t) - \frac{1}{\mu^n}\frac{1}{\sqrt{n}} - \frac{1}{\mu^n}\frac{1}{\sqrt{n}}\mathcal{A}^n(t) \\ &\quad + \frac{1}{\mu^n}\left(\tilde{V}_O^n\left(\frac{S_O^n(T_O^n(t))}{n} + \frac{Q_O^n(t)}{n} - \frac{\mathcal{A}^n(t)}{n}\right) - \tilde{V}_O^n\left(\frac{S_O^n(T_O^n(t))}{n} + \frac{1}{n}\right)\right). \end{aligned} \tag{31}$$

We will first show that $\sqrt{n}U_O^n - \sqrt{n}L_O^n \Rightarrow 0$ as $n \to \infty$, and then show

$$\sqrt{n}L_O^n \Rightarrow \frac{(1-\alpha)w_I}{(1-\alpha)w_I + \alpha w_O}\frac{Z}{\mu} \tag{32}$$

as $n \to \infty$. The inequality (29) and the converging together lemma then establish (28), and so the weak convergence in (14) follows.

Since

$$\frac{1}{\sqrt{n}}\mathcal{A}^n(t) = \tilde{N}^n\left(\overline{\tau}^n\left(t + \frac{W_O^n(t)}{1-\alpha}\right)\right) - \tilde{N}^n\left(\overline{\tau}^n(t)\right) + \int_t^{t+(1-\alpha)^{-1}W_O^n(t)} \gamma\tilde{Q}_O^n(s)ds,$$

and Lemma 1 establishes $\overline{\tau}^n \to 0$ and $W_O^n \to 0$ a.s., u.o.c., it follows from the functional central limit theorem, continuous mapping theorem, and the weak convergence of $\tilde{Q}_O^n$ in (13) that

$$\frac{1}{\sqrt{n}}\mathcal{A}^n \Rightarrow 0 \tag{33}$$

as $n \to \infty$. It follows from Lemma 3 in Iglehart and Whitt (1970) that for any $t > 0$ $n^{-1/2}v_{S_O^n(T_O^n(t))+1} \to 0$ in probability, as $n \to \infty$. Now, the sequence $\{T_O^n\}$ is tight in D

because $|T_O^n(t) - T_O^n(s)| \leq |t-s|$. On any subsequence $\{n_k\}$ on which

$$T_O^{n_k} \Rightarrow T_O$$

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

$$\frac{S_O^{n_k} \circ T_O^{n_k}}{n_k} \Rightarrow \mu T_O$$

as $n_k \to \infty$. Furthermore, on this same subsequence, by the convergences in (33) and Lemma 1, $n_k^{-1}\mathcal{A}^{n_k} \Rightarrow 0$ and $n_k^{-1}Q_O^{n_k} \to 0$ a.s., u.o.c. as $n_k \to \infty$. Therefore, because by Donsker's theorem $\tilde{V}_O^n$ weakly converges to a continuous limit process,

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

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

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

as $n \to \infty$. We conclude from (30) that as $n \to \infty$

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

We now establish the weak convergence in (32). An argument similar to that in the above paragraph shows

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

as $n \to \infty$. Hence, the representation of $\sqrt{n}L_O^n$ in (31), Theorems 1 and 2 (specifically, the resulting convergence in (13)), the convergence in (33), and the continuous mapping theorem establish (32).

Proof of (15)

First observe that it is sufficient to show that for any $t > 0$,

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

as $n \to \infty$. Now, $\frac{d}{dt}T_I^n(t) = \alpha$ and $\frac{d}{dt}T_O^n(t) = 1-\alpha$ if and only if $Q_I^n(t) > 0$ and $Q_O^n(t) > 0$. Hence it is enough to show

$$\begin{aligned} P\left(Q_I^n(t) > 0\right) = P\left(\tilde{Q}_I^n(t) > 0\right) &\to 1 \\ P\left(Q_O^n(t) > 0\right) = P\left(\tilde{Q}_O^n(t) > 0\right) &\to 1 \end{aligned}$$

as $n \to \infty$, which follows from the weak convergence in (13). □

Proof of Proposition 1

We must show the following.

(i) For any $T > 0$, $\sup_{0\le t\le T}\left|\frac{w_I}{\alpha}\tilde{Q}_I^n(t) - \frac{w_O}{1-\alpha}\tilde{Q}_O^n(t)\right| \to 0$, in probability, as $n \to \infty$.

(ii) As $n \to \infty$, $\left(\tilde{Q}^n, \tilde{I}^n\right) \Rightarrow (Z, L)$.

(iii) As $n \to \infty$, $\tilde{W}_I^n \Rightarrow \frac{\alpha w_O}{(1-\alpha)w_I + \alpha w_O}\frac{Z}{\mu}$ and $\tilde{W}_O^n \Rightarrow \frac{(1-\alpha)w_I}{(1-\alpha)w_I + \alpha w_O}\frac{Z}{\mu}$.

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

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

With ξ_n and ξ_n^* defined exactly as in the proof of Theorem 1, observe that when

$$\frac{w_I}{\alpha}\tilde{Q}_I^n\left(\xi_n^*\right) > (\le)\frac{w_O}{1-\alpha}\tilde{Q}_O^n\left(\xi_n^*\right),$$

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

$$\begin{aligned} S_I^n(t) - S_I^n\left(\xi_n^*-\right) &\ge (\le) \left\lfloor\frac{t-\xi_n^*-}{l^n}\right\rfloor \lfloor\alpha n^{1/3}\mu^n\rfloor \\ S_O^n(t) - S_O^n\left(\xi_n^*-\right) &\le (\ge) \left\lfloor\frac{t-\xi_n^*-}{l^n}\right\rfloor \lceil(1-\alpha)n^{1/3}\mu^n\rceil. \end{aligned}$$

Then, substitution of the above bounds into the equivalent of (25) in the proof of Theorem 1 in this setting (specifically, replace $S_I^n(T_I^n(t)) - S_I^n(T_I^n(\xi_n^*-))$ with $S_I^n(t) - S_I^n(\xi_n^*-)$ and $S_O^n(T_O^n(t)) - S_O^n(T_O^n(\xi_n^*-))$ with $S_O^n(t) - S_O^n(\xi_n^*-)$) shows

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

Noting that $(nl^n)^{-1} = n^{-1/3} \to \infty$ as $n \to \infty$, the remainder of the proof proceeds exactly as the proof of Theorem 1.

(ii): We first observe that the number-in-system process can be equivalently written as

$$Q_I^n(t) + Q_O^n(t) = A^n(t) - N\left(\int_0^t \gamma Q_O^n(s)ds\right) - \left\lfloor \frac{t}{l^n} \right\rfloor n^{1/3}\mu^n + I^n(t), \tag{34}$$

where I^n is a non-decreasing process for which $\int_0^\infty \left(Q_I^n(t) + Q_O^n(t)\right) dI^n(t) = 0$ and $I^n(0) = 0$. Specifically, the process I^n may increase only at discrete review time points $\{l^n, 2l^n, 3l^n, \ldots\}$, and is defined recursively as

$$\begin{aligned} I^n(0) &= 0 \\ I^n(il^n) &= I^n((i-1)l^n) + \left[n^{1/3}\mu^n - Q_I^n(il^n-) - Q_O^n(il^n-)\right]^+. \end{aligned}$$

The process I^n tracks the cumulative amount of spare capacity. To see the equation (34) holds, note that

$$Q_I^n(t) + Q_O^n(t) = A^n(t) - N\left(\int_0^t \gamma Q_O^n(s)ds\right) - S_I^n\left(\left\lfloor \frac{t}{l^n} \right\rfloor l^n\right) - S_O^n\left(\left\lfloor \frac{t}{l^n} \right\rfloor l^n\right),$$

and, for every $i \in \{0, 1, \ldots\}$,

$$\begin{aligned} &S_I^n(il^n) + S_O^n(il^n) - S_I^n((i-1)l^n) - S_O^n((i-1)l^n) \\ &= n^{1/3}\mu^n \mathbf{1}\{Q_I^n(il^n-) + Q_O^n(il^n-) \geq n^{1/3}\mu^n\} \\ &\quad + \left(Q_I^n(il^n-) + Q_O^n(il^n-)\right)\mathbf{1}\{Q_I^n(il^n-) + Q_O^n(il^n-) < n^{1/3}\mu^n\}. \end{aligned}$$

Finally,

$$\int_0^\infty \left(Q_I^n(t) + Q_O^n(t)\right) dI^n(t) = \sum_{i=0}^\infty \left(Q_I^n(il^n) + Q_O^n(il^n)\right)\left(I^n(il^n) - I^n((i-1)l^n)\right) = 0.$$

It follows from (34) that

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

for

$$\begin{aligned}
\tilde{X}^n(t) &= \tilde{A}^n(t) - \tilde{N}^n\left(\overline{\tau}^n(t)\right) + \sqrt{n}t\left(\lambda^n - \left\lfloor \frac{t}{l^n} \right\rfloor \left(\frac{l^n}{t}\right)\mu^n\right) \\
\tilde{\epsilon}^n(t) &= \gamma \int_0^t \left(\frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I}\tilde{Q}^n(s) - \tilde{Q}_O^n(s)\right) ds.
\end{aligned}$$

The properties of I^n then imply

$$\left(\tilde{Q}^n, \tilde{I}^n\right) = (\phi^\kappa, \psi^\kappa)\left(\tilde{X}^n + \epsilon^n\right).$$

The functional central limit theorem, the fact that $\tilde{N}^n \circ \overline{\tau}^n \Rightarrow 0$ as $n \to \infty$ (by the same argument as that directly following (26) in the proof of Theorem 1), the heavy traffic assumption (8), the state space collapse in part (i), and the representation $(Z, L) = (\phi^\kappa, \psi^\kappa)(e + \sigma B)$ in (12) then establish

$$(\phi^\kappa, \psi^\kappa)\left(\tilde{X}^n + \tilde{\epsilon}^n\right) \Rightarrow (\phi^\kappa, \psi^\kappa)(\sigma B + \theta e) = (Z, L)$$

as $n \to \infty$.

(iii): We show that the weak convergence in (14) remains valid; the argument showing (15) holds is exactly as in the proof of Theorem 3. The number of batches required to serve all customers in the inline queue exceeds $\lfloor Q_I^n(t)/(n^{1/3}\mu^n) \rfloor$ and is less than $\lceil Q_I^n(t)/n^{1/3}\mu^n \rceil$. Since each batch requires l^n time units to process

$$l^n \left\lfloor \frac{Q_I^n(t)}{n^{1/3}\mu^n} \right\rfloor \leq W_I^n(t) \leq \left\lceil \frac{Q_I^n(t)}{n^{1/3}\mu^n} \right\rceil,$$

and so

$$0 \leq \sqrt{n}W_I^n(t) - \sqrt{n}l^n \left\lfloor \frac{Q_I^n(t)}{n^{1/3}\mu^n} \right\rfloor \leq \sqrt{n}l^n.$$

Since $\sqrt{n}l^n \to 0$ as $n \to \infty$ and by parts (i) and (ii) of this Proposition the weak convergence in (13) remains valid,

$$\sqrt{n}l^n \frac{Q_I^n}{n^{1/3}\mu^n} = \frac{1}{\mu^n}\tilde{Q}_I^n \Rightarrow \frac{\alpha w_O}{(1-\alpha)w_I + \alpha w_O}\frac{Z}{\mu}.$$

We conclude

$$\tilde{W}_I^n \Rightarrow \frac{\alpha w_O}{(1-\alpha)w_I + \alpha w_O} \frac{Z}{\mu}$$

as $n \to \infty$.

Since whenever the number of customers in the offline queue exceeds $(1-\alpha)n^{1/3}\mu^n$ at a discrete review time point, at least $(1-\alpha)n^{1/3}\mu^n$ customers are served,

$$\left(\frac{Q_O^n(t)}{(1-\alpha)n^{1/3}\mu^n} + 1 \right) l^n$$

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

$$\mathcal{A}^n(t) \equiv N\left(\int_0^{t+\left(\frac{Q_O^n(t)}{(1-\alpha)n^{1/3}\mu^n}+1\right)l^n} \gamma Q_O^n(s)ds \right) - N\left(\int_0^t \gamma Q_O^n(s)ds \right).$$

Therefore,

$$l_n \left\lfloor \frac{Q_O^n(t) - \mathcal{A}^n(t)}{n^{1/3}\mu^n} \right\rfloor \leq W_O^n(t) \leq l^n \left\lceil \frac{Q_O^n(t)}{n^{1/3}\mu^n} \right\rceil.$$

It follows from the observation that

$$\left(\frac{Q_O^n(t)}{(1-\alpha)n^{1/3}\mu^n} + 1 \right) l^n = \frac{Q_O^n(t)}{(1-\alpha)\mu^n n} + l^n \to 0$$

as $n \to \infty$ that

$$\mathcal{A}^n \Rightarrow 0$$

as $n \to \infty$ by identical argument as that in the proof of Theorem 3. As in the preceding paragraph, we conclude

$$\tilde{W}_O^n \Rightarrow \frac{(1-\alpha)w_I}{(1-\alpha)w_I + \alpha w_O} \frac{Z}{\mu}$$

as $n \to \infty$. □

Proofs of Lemma 1

Define

$$\overline{X}^n(t) \equiv \overline{A}^n(t) - \overline{S}_I^n\left(T_I^n(t)\right) - \overline{S}_O^n\left(T_O^n(t)\right) - \overline{N}^n\left(\bar{\tau}^n(t)\right) + \left(\lambda^n - \mu^n\right)t.$$

Then, for all $t \geq 0$,

$$\bar{Q}^n(t) = \bar{X}^n(t) - \bar{\tau}^n(t) + \mu^n I^n(t).$$

Since I^n is non-decreasing, $I^n(0) = 0$ and $\int_0^\infty \overline{Q}^n(t) d\left(\mu^n I^n(t)\right) = 0$, the process $\left(\overline{Q}^n, \mu^n I^n\right)$ can be represented in terms of the conventional two-sided regulator mapping as follows

$$\left(\overline{Q}^n, \mu^n I^n\right) = (\phi, \psi)\left(\overline{X}^n - \overline{\tau}^n\right).$$

Since $\overline{\tau}^n$ is a non-decreasing process, Lemma 5.1 in Kruk et al. (2006) establishes

$$\phi\left(\overline{X}^n - \overline{\tau}^n\right) \leq \phi\left(\overline{X}^n\right).$$

The functional strong law of large numbers and the heavy traffic assumption (8) establish

$$\overline{X}^n \to 0 \text{ a.s., u.o.c.,}$$

as $n \to \infty$, which implies, because ϕ is a continuous function, that

$$\phi\left(\overline{X}^n\right) \to 0 \text{ a.s., u.o.c..}$$

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

$$\overline{Q}^n \to 0 \text{ a.s., u.o.c.,}$$

as $n \to \infty$. It then follows that for any $T > 0$,

$$\sup_{0 \leq t \leq T} |\overline{\tau}^n(t)| = \int_0^T \gamma \overline{Q}^n(s) ds \to 0,$$

as $n \to \infty$, and so

$$\overline{\tau}^n \to 0 \text{ a.s., u.o.c.,}$$

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

$$I^n = \frac{1}{\mu^n} \psi\left(\overline{X}^n - \overline{\tau}^n\right) \to 0 \text{ a.s., u.o.c.,}$$

as $n \to \infty$. The condition (6) then implies

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

as $n \to \infty$.

It remains to show

$$W_O^n \to 0 \text{ a.s., u.o.c.,}$$

as $n \to \infty$. First recall that for

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

defined as in the proof of Theorem 3,

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

Define

$$\overline{V}_O^n(t) \equiv \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} \left(v_i^O - 1\right),$$

and observe that

$$U_O^n(t) = \frac{1}{\mu^n} \left(\overline{V}_O^n \left(\frac{1}{n} S_O^n \left(T_O^n(t)\right) + \overline{Q}_O^n(t) \right) - \overline{V}_O^n \left(\frac{1}{n} S_O^n \left(T_O^n(t)\right) \right) \right) + \frac{1}{\mu^n} \overline{Q}_O^n(t).$$

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

$$\overline{Q}_O^n \to 0 \text{ a.s., u.o.c.,}$$

as $n \to \infty$. Therefore, because also $\overline{V}_O^n \to 0$ a.s., u.o.c. as $n \to \infty$, it follows that $\overline{U}_O^n \to 0$ a.s., u.o.c. as $n \to \infty$, we conclude

$$W_O^n \to 0 \text{ a.s., u.o.c.,}$$

as $n \to \infty$. □

Proof of Lemma 2

Fix $T > 0$ and $\epsilon > 0$. Define

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

Then,

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

Since $\tilde{I}^n$ is non-decreasing, $\tilde{I}^n(0) = 0$, and the condition (7) implies $\int_0^\infty \tilde{Q}^n(t) d\left(\mu^n \tilde{I}^n(t)\right) = 0$, the process $\left(\tilde{Q}^n, \mu^n \tilde{I}^n\right)$ can be represented in terms of the conventional two-sided regulator mapping as follows

$$\left(\tilde{Q}^n, \mu^n \tilde{I}^n\right) = (\phi, \psi)\left(\tilde{\chi}^n - \tilde{\mathcal{A}}^n\right). \tag{35}$$

Since $\tilde{\mathcal{A}}^n$ is a non-decreasing process, Lemma 5.1 in Kruk et al. (2006) establishes that

$$\phi\left(\tilde{\chi}^n - \tilde{\mathcal{A}}^n\right)(t) \leq \phi\left(\tilde{\chi}^n\right)(t) \text{ for all } t \geq 0. \tag{36}$$

The functional central limit theorem, continuous mapping theorem, and heavy traffic assumption (8) establish

$$\phi\left(\tilde{\chi}^n\right) \Rightarrow \phi\left(\theta e + \sigma W\right),$$

as $n \to \infty$. Since weak convergence implies the random variable $\sup_{0 \leq t \leq T} \phi\left(\tilde{\chi}^n\right)(t)$ is tight, there exists B and n_0 large enough so that

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

Therefore, it follows from the representation (35) and the upper bound (36) that

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

□

Acknowledgments

We would like to thank Mor Armony for helpful discussions related to 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.

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 encountersForthcoming in the *Manufacuring & Service Operations Management*.

Bookbinder, J. H., Noor, A. I., 1985. Setting job-shop due dates with service level constaints. *J. Oper. Res. Soc.* **36**, 1017–1026.

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.

Dai, J. G., Dai, W., 1999. A heavy traffic limit theorem for a class of open queueing networks with finite buffers. *Queueing Systems* **32**, 5–40.

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.

Extend: Professional simulation tools, 2003. Version 6.

Flatto, L., McLean, H. P., 1977. Two queue in parallel. *Communications in Pure and Applied Mathematics* **30**, 255–263.

Foschini, G. J., Salz, J., 1978. A basic dynamic routing problem and diffusion. *IEEE Trans. Commun.* **26**, 320–327.

Gamarnik, D., Zeevi, A., 2006. Validity of heavy traffic steady-state approximations in generalized jackson networks. *Annals of Applied Probability* **16**, 56–90.

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

Hopp, W. J., Sturgis, M. R., 2001. A simple, robust leadtime-quoting policy. *Manufacturing & Service Operations Management* **3**, 321–336.

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.

Keskinocak, P., Ravi, R., Tayur, S., 2001. Scheduling and reliable lead time quotation for orders with availability intervals and lead time sensitive revenues. *Management Science* **47**, 264–279.

Kruk, L., Lehoczky, J., Ramanan, K., Shreve, S., 2005. An explicit formula for double reflected processes in [0,a]. Submitted.

Maister, D., 1985. The psychology of waiting in lines. Lexington Books, Lexington, MA.

McDonald, D. R., 1996. Overloading parallel servers when arrivals join the shortest queue. Springer-Verlag, NY, lecture Notes in Statistics 117.

Munichor, N., Rafaeli, A., 2007. Number of apologies? Customer reactions to telephone waiting time fillers. *Journal of Applied Probability* **92**, 511–518.

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.

Skorokhod, A. V., 1961. Stochastic equations for diffusions in a bounded region. *Theor. of Prob. and Its Appl.* **6**, 264–274.

Spearmen, M. L., Zhang, R. Q., 1999. Optimal leadtimes. *Management Science* **45**, 290–295.

Taylor, S., 1994. Waiting for service: the realationship between delays and evaluations of service. *Journal of Marketing* **58**, 56–69.

Tatsu Ride Statistics, 2007. http://en.wikipedia.org/wiki/Tatsu.

Turner, S. R. E., 2000. Large deviations for join the shorter queue. American Mathematical Society, Province, RI.

Ward, A. R., Glynn, P. W., 2003. Properties of the reflected Ornstein-Uhlenbeck process. *Queueing Systems* **44**, 109–123.

Ward, A. R., Kumar, S., 2007. Asymptotically optimal control of a queue with impatient customers. Forthcoming in *Mathematics of Operations Research*.

Wein, L. M., 1991. Due date setting and priority sequencing in a multiclass M/G/1 queue. *Management Science* **37**, 834–850.

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

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

Whitt, W., 2002. Stochastic-Process Limits. Springer, New York.