

A Single-Server Queue with a Delayed Service Option and Exponential Abandonment Times

Vasiliki Kostami, Sriram Dasu, and Amy Ward

August 26, 2007

Abstract

This paper ...

1 Introduction

[TO BE WRITTEN, ALONG WITH LITERATURE REVIEW.][ALSO REMEMBER TO PUT IN NUMERIC RESULTS THAT SHOW THE AFFECT OF CUSTOMER ABANDONMENTS FROM THE DELAYED SERVICE QUEUE.]

We consider a single-server queueing model in which customers may choose between waiting for service in real-time, and returning for service at a dynamically specified future time point, as shown in Figure 1. Customers waiting for service in real-time are physically present in the queue, and will wait as long as necessary to obtain their service. Customers that choose the delayed service option need not be physically present, and may not return for service.

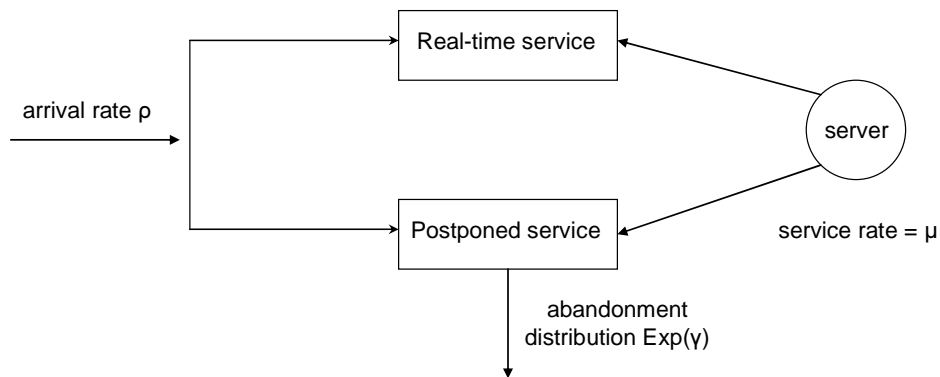


Figure 1: The model

The service discipline is generalized processor sharing. Specifically, when

there are both customers waiting for real-time service and delayed service, the server processes the customers in the real-time queue at rate $\mu\alpha$ and those in the delayed service queue at rate $\mu(1 - \alpha)$.

2 Model Formulation

Our model formulation consists of a sequence of generalized join-the-shorter queue queueing systems in which customer demand becomes large and service is rendered quickly, so that the server utilization approaches unity. We begin by presenting the basic system model in Subsection 2.1. We then describe our heavy traffic asymptotic regime in Subsection 2.2.

2.1 Basic System Model

Let w_R and w_D be the waiting costs per time unit for the real-time and the delayed service queues respectively. Suppose that at time $t \geq 0$, the length of the real-time service queue is $Q_R(t)$ and the delayed service queue is $Q_D(t)$. The expected total amount of processing required by all the customers in the real-time and delayed service queues is $\mu^{-1}Q_R(t)$ and $\mu^{-1}Q_D(t)$ respectively. Assuming the system is heavily utilized, so that the time either the real-time or delayed service queue is empty is small, the real-time service queue is processed at rate α , and the delayed service queue is processed at rate $(1 - \alpha)$. Hence a customer desiring to minimize his cost of waiting joins the real-time service queue if

$$w_R \frac{Q_R(t)}{\mu\alpha} \leq w_D \frac{Q_D(t)}{\mu(1 - \alpha)},$$

and otherwise joins the delayed service queue.

Not all of the customers in the delayed service queue return for processing. Therefore, $[\mu(1 - \alpha)]^{-1}Q_D(t)$ overestimates the time a customer joining the delayed service queue at time t must wait for service. However, we will show that in our heavy traffic asymptotic regime such an overestimation is small, because the probability any individual customer abandons the delayed service queue becomes small. See Theorems 2 and 3 for theoretical support of this statement and our simulation results in Section 5 for numeric support. (Of course, the presence of customer abandonments from the delayed

service queue does affect the approximations we develop for the queue-length processes.)

Let A , S_R , and S_D be independent renewal processes having rates λ , μ , and μ respectively. $A(t)$ denotes the cumulative number of arrivals to the system in $[0, t]$. $S_R(t)$ and $S_D(t)$ denote respectively the cumulative number of departures from the real-time and delayed service queues after the server has devoted t units of time to the queue working at rate 1. Let R be an independent, standard Poisson process. We use R to track the cumulative number of customers that renege from the delayed service queue. The evolution equations for Q_R and Q_D are

$$Q_R(t) \equiv \sum_{i=1}^{A(t)} \mathbf{1}\left\{\frac{w_R}{\mu\alpha} Q_R(t_{i-}) \leq \frac{w_D}{\mu(1-\alpha)} Q_D(t_{i-})\right\} - S_R(T_R(t)) \quad (2.1)$$

$$Q_D(t) \equiv \sum_{i=1}^{A(t)} \mathbf{1}\left\{\frac{w_D}{\mu\alpha} Q_R(t_{i-}) > \frac{w_D}{\mu(1-\alpha)} Q_D(t_{i-})\right\} - R\left(\int_0^t \gamma Q_D(s) ds\right) - S_D(T_D(t)), \quad (2.2)$$

where

$$T_R(t) \equiv \int_0^t \frac{\alpha \mathbf{1}\{Q_R(s) > 0\}}{\alpha \mathbf{1}\{Q_R(s) > 0\} + (1-\alpha) \mathbf{1}\{Q_D(s) > 0\}} ds \quad (2.3)$$

$$T_D(t) \equiv \int_0^t \frac{(1-\alpha) \mathbf{1}\{Q_D(s) > 0\}}{\alpha \mathbf{1}\{Q_R(s) > 0\} + (1-\alpha) \mathbf{1}\{Q_D(s) > 0\}} ds. \quad (2.4)$$

Define

$$Q \equiv Q_R + Q_D.$$

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

$$I(t) \equiv \int_0^t \mathbf{1}\{Q(s) = 0\} ds \quad (2.5)$$

is the cumulative server idletime. Then,

$$T_R(t) + T_D(t) + I(t) = t \quad (2.6)$$

$$\int_0^\infty Q(t) dI(t) = 0. \quad (2.7)$$

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 reneging rate γ , the server-sharing constant α , and the waiting costs w_R and w_D all remain constant. Our convention is to superscript any process or quantity associated with the n th system by n .

Let $\{u_i, i \geq 1\}$, $\{v_i^D, i \geq 1\}$ and $\{v_i^R, i \geq 1\}$ be three independent, i.i.d. sequences of non-negative, mean 1 random variables having finite variance. Further assume that $\{v_i^D, i \geq 1\}$ and $\{v_i^R, i \geq 1\}$ 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 real-time service queue after the server has worked at rate 1 for t time units is

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

and from the delayed service queue is

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

Define the fluid scaled quantities

$$\begin{aligned}
\bar{A}^n(t) &\equiv \frac{1}{n}A^n(t) - \lambda^n t \\
\bar{S}_R^n(t) &\equiv \frac{1}{n}S_R^n(t) - \mu^n t \\
\bar{S}_D^n(t) &\equiv \frac{1}{n}S_D^n(t) - \mu^n t \\
\bar{R}^n(t) &\equiv \frac{1}{n}R(nt) - t \\
\bar{Q}_R^n(t) &\equiv \frac{1}{n}Q_R^n(t) \\
\bar{Q}_D^n(t) &\equiv \frac{1}{n}Q_D^n(t) \\
\bar{Q}^n(t) &\equiv \frac{1}{n}Q^n(t) \\
\bar{\tau}^n(t) &\equiv \frac{1}{n} \int_0^t \gamma Q_D^n(s) ds,
\end{aligned}$$

and the diffusion scaled quantities

$$\begin{aligned}
\tilde{A}^n(t) &\equiv \sqrt{n} \left(\frac{1}{n} A^n(t) - \lambda^n t \right) \\
\tilde{Q}_R^n(t) &\equiv \frac{1}{\sqrt{n}} Q_R^n(t) \\
\tilde{Q}_D^n(t) &\equiv \frac{1}{\sqrt{n}} Q_D^n(t) \\
\tilde{S}_R^n(t) &\equiv \sqrt{n} \left(\frac{1}{n} S_R^n(t) - \mu^n t \right) \\
\tilde{S}_D^n(t) &\equiv \sqrt{n} \left(\frac{1}{n} S_D^n(t) - \mu^n t \right) \\
\tilde{I}^n(t) &\equiv \sqrt{n} I^n(t) \\
\tilde{R}^n(t) &\equiv \sqrt{n} \left(\frac{1}{n} R(nt) - t \right)
\end{aligned}$$

As n increases,

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

where $\mu \in \mathfrak{R}$. (Note the slight abuse of notation because μ is no longer the service rate introduced in Section 2.1. The service rate in the n th system is

$n\mu^n$.) Furthermore, λ^n and μ^n become close at rate \sqrt{n} ; specifically,

$$\sqrt{n}(\lambda^n - \mu^n) \rightarrow \theta, \quad (2.8)$$

as $n \rightarrow \infty$, where $\theta \in \mathfrak{R}$.

The functional strong law of large numbers establishes

$$(\bar{A}^n, \bar{S}_R^n, \bar{S}_D^n) \rightarrow (0, 0, 0) \text{ u.o.c., a.s.}, \quad (2.9)$$

as $n \rightarrow \infty$. 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 (2.9) 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}_R^n , and \tilde{S}_D^n can be approximated by Brownian motion. For this, 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 $D([0, \infty), \mathfrak{R}^d)$ be the space of right continuous functions with left limits (RCLL) in \mathfrak{R}^d having time domain $[0, \infty)$. We endow $D([0, \infty), \mathfrak{R}^d)$ 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 $(D([0, \infty), \mathfrak{R}^d), M^d)$ 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 $(D([0, \infty), \mathfrak{R}^d), M^d)$ converge weakly to the probability measure on $(D([0, \infty), \mathfrak{R}^d), M^d)$ induced by the stochastic process ξ . Note that we suppress d from the notation unless necessary.

Let B_u , B_{v_R} and B_{v_D} be independent, standard Brownian motions. The functional central limit theorem and the assumed independence of the inter-arrival and service time sequences establish

$$\left(\tilde{A}^n, \tilde{S}_R^n, \tilde{S}_D^n\right) \Rightarrow \left(\sqrt{\lambda \text{var}(u_1)} B_u, \sqrt{\mu \text{var}(v_1)} B_{v_R}, \sqrt{\mu \text{var}(v_1)} B_{v_D}\right). \quad (2.10)$$

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 [1] or Whitt [9]. We also often use the notation e to denote the identity process

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

3 Asymptotic Analysis

We show that the two-dimensional queue-length process described in our basic system model reduces to one dimension in our heavy traffic asymptotic regime, and identify the one-dimensional limit process. We state our results in Subsection 3.1, and prove them in Subsection 3.2.

3.1 Convergence Results

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

Theorem 1 *For any $T > 0$, $\sup_{0 \leq t \leq T} |\frac{w_R}{\alpha} \tilde{Q}_R^n(t) - \frac{w_D}{1-\alpha} \tilde{Q}_D^n(t)| \rightarrow 0$ in probability as $n \rightarrow \infty$.*

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

$$\sigma^2 = \lambda \text{var}(u_1) + \mu \text{var}(v_1^R).$$

Define Z as the strong solution to the stochastic equation

$$Z(t) = \theta t - \gamma \frac{(1-\alpha)w_R}{\alpha w_D + (1-\alpha)w_R} \int_0^t Z(s) ds + \sigma B(t) + L(t) \geq 0, \quad t \geq 0, \quad (3.1)$$

where L is non-decreasing, $L(0) = 0$ and $\int_0^\infty Z(t) dL(t) = 0$. The existence of a strong solution to (3.1) 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 $x \in D([0, \infty), \mathfrak{R})$ 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) $z(t) = x(t) - \kappa \int_0^t z(s)ds + l(t) \in [0, \infty)$ for all $t \geq 0$;

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

Specifically, for

$$\kappa \equiv \gamma \frac{(1 - \alpha)w_R}{\alpha w_D + (1 - \alpha)w_R},$$

it follows that

$$(Z, L) = (\phi^\kappa, \psi^\kappa)(e + \sigma B). \quad (3.2)$$

Proposition 3 part (i) in Reed and Ward [4] establishes the existence and uniqueness of the regulator mapping in Definition 1¹, and so the representation (3.2) guarantees that there is a unique strong solution to the stochastic equation in (3.1). 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 [7].

Our next theorem establishes that the process Z in (3.1) approximates the total number of customers in either the real-time or delayed service queues.

Theorem 2 As $n \rightarrow \infty$,

$$\tilde{Q}_R^n + \tilde{Q}_D^n \Rightarrow Z.$$

Together, Theorems 1 and 2 imply a separate approximation for the number of customers in the real-time queue, and for the number of customers in the delayed service queue. In particular, we observe that

$$\tilde{Q}_R^n \Rightarrow \frac{\alpha w_D}{(1 - \alpha)w_R + \alpha w_D} Z \text{ and } \tilde{Q}_D^n \Rightarrow \frac{(1 - \alpha)w_R}{(1 - \alpha)w_R + \alpha w_D} Z \quad (3.3)$$

as $n \rightarrow \infty$.

Our basic system model presented in Section 2.1 relies on the assumption that the amount of time a customer joining either the real-time or delayed service queue (assuming he does not abandon) would wait to receive service

¹Actually, the regulator mapping in Definition 1 is a specific instance of the more general regulator mapping in [4].

at time t can be approximated from the queue-length processes. It is not obvious that the queue-length of the delayed service queue can be used to estimate waiting times because the number of customers that will abandon the delayed service queue is not known. However, our next theorem shows that such an approximation is possible in our heavy traffic asymptotic regime.

Let W_R^n and W_D^n be the workload processes in the real-time and delayed service 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 processing occurs at rate 1. Define $\tilde{W}_R^n = \sqrt{n}W_R^n$ and $\tilde{W}_D^n = \sqrt{n}W_D^n$.

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

$$\tilde{W}_R^n \Rightarrow \frac{\alpha w_D}{(1 - \alpha)w_R + \alpha w_D} \frac{Z}{\mu} \text{ and } \tilde{W}_D^n \Rightarrow \frac{(1 - \alpha)w_R}{(1 - \alpha)w_R + \alpha w_D} \frac{Z}{\mu}.$$

The weak convergence in (3.3) resulting from Theorems 1 and 2, and Theorem 3, establish that

$$\sqrt{n}W_R^n \approx \frac{Q_R^n}{\sqrt{n}\mu^n} \text{ and } \sqrt{n}W_D^n \approx \frac{Q_D^n}{\sqrt{n}\mu^n}$$

Because in heavy traffic both the real-time and delayed service queues are usually non-empty, the real-time service queue generally receives proportion α of the server processing power, and the delayed service queue generally receives proportion $(1 - \alpha)$. Therefore, a reasonable approximation of the time a customer joining the real-time service queue at time t would have to wait to receive service is

$$\frac{Q_R^n}{n\mu^n\alpha} \approx \frac{\alpha w_D}{(1 - \alpha)w_R + \alpha w_D} \frac{Z}{\sqrt{n}\mu^n\alpha},$$

and is

$$\frac{Q_D^n}{n\mu^n\alpha} \approx \frac{(1 - \alpha)w_R}{(1 - \alpha)w_R + \alpha w_D} \frac{Z}{\sqrt{n}\mu^n\alpha}$$

for a customer joining the delayed service queue (that does not abandon). We provide numeric results in Section 5 supporting this proposed approximation.

3.2 Proofs of Convergence Results

We require the following two Lemmas, whose proofs can be found in the appendix.

Lemma 1 *Let $W^n = W_R^n + W_D^n$. As $n \rightarrow \infty$,*

$$(\bar{Q}^n, \bar{W}^n, \bar{\tau}^n, \bar{T}_R^n + \bar{T}_D^n, \bar{I}^n) \rightarrow (0, 0, 0, e, 0).$$

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

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

for all $n \geq n_0$.

3.2.1 Proof of Theorem 1

The structure of our proof follows the proof of Theorem 1 in Section 5 in Reiman [6], 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 $\epsilon > 0$,

$$P \left(\sup_{0 \leq t \leq T} \left| \frac{w_R}{\alpha} \tilde{Q}_R^n(t) - \frac{w_D}{1-\alpha} \tilde{Q}_D^n(t) \right| > \epsilon \right) \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (3.4)$$

Fix $\epsilon > 0$ and let

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

It will also be useful to define the processes

$$\begin{aligned}
\tilde{U}_I^n(t, s, u, v) &\equiv -\frac{w_R}{\alpha} \left\{ \tilde{S}_R^n(u + \alpha(t - s)) - \tilde{S}_R^n(u) \right\} \\
&\quad + \frac{w_D}{1 - \alpha} \left\{ \tilde{S}_D^n(v + (1 - \alpha)(t - s)) - \tilde{S}_D^n(v) \right\} \\
&\quad - \frac{w_D}{1 - \alpha} \left\{ \tilde{A}^n(t) - \tilde{A}^n(s) \right\} \\
&\quad + \left\{ \frac{w_D}{1 - \alpha} (\mu^n - \lambda^n) - \mu^n \left(w_R + \frac{\alpha}{1 - \alpha} w_D \right) \right\} \sqrt{n}(t - s) \\
\tilde{U}_{II}^n(t, s, u, v) &\equiv -\frac{w_D}{1 - \alpha} \left\{ \tilde{S}_D^n(v + (1 - \alpha)(t - s)) - \tilde{S}_D^n(v) \right\} \\
&\quad + \frac{w_R}{\alpha} \left\{ \tilde{S}_R^n(u + \alpha(t - s)) - \tilde{S}_R^n(u) \right\} \\
&\quad - \frac{w_R}{\alpha} \left\{ \tilde{A}^n(t) - \tilde{A}^n(s) \right\} \\
&\quad + \left\{ \frac{w_R}{\alpha} (\mu^n - \lambda^n) - \mu^n \left(\frac{1 - \alpha}{\alpha} w_R + w_D \right) \right\} \sqrt{n}(t - s).
\end{aligned}$$

An upper bound for the left-hand-side of (3.4)

First assume $(w_R/\alpha)\tilde{Q}_R^n(\xi_n^*) > (w_D/(1 - \alpha))\tilde{Q}_D^n(\xi_n^*)$. Then, for $\xi_n^* \leq t \leq \xi_n$, all customers join the delayed service queue, and so

$$\begin{aligned}
&\left| \frac{w_R}{\alpha} \tilde{Q}_R^n(t) - \frac{w_D}{1 - \alpha} \tilde{Q}_D^n(t) \right| \\
&= \frac{w_R}{\alpha} \tilde{Q}_R^n(\xi_n^* -) - \frac{w_D}{1 - \alpha} \tilde{Q}_D^n(\xi_n^* -) - \frac{w_R}{\alpha} \frac{1}{\sqrt{n}} \{S_R^n(T_R^n(t)) - S_R^n(T_R^n(\xi_n^* -))\} \\
&\quad + \frac{w_D}{1 - \alpha} \frac{1}{\sqrt{n}} \{S_D^n(T_D^n(t)) - S_D^n(T_D^n(\xi_n^* -))\} + \frac{w_D}{1 - \alpha} \frac{1}{\sqrt{n}} R^n \left(\int_{\xi_n^* -}^t \gamma Q_D^n(s) ds \right) \\
&\quad - \frac{w_D}{1 - \alpha} \left\{ \tilde{A}^n(t) - \tilde{A}^n(\xi_n^* -) + \sqrt{n} \lambda^n (t - \xi_n^*) \right\}. \tag{3.5}
\end{aligned}$$

The real-time service queue does not empty during $[\xi_n^*, \xi_n]$, so that

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

The delayed service queue may empty during $[\xi_n^*, \xi_n]$, so that

$$T_D^n(t) - T_D^n(\xi_n^* -) \leq (1 - \alpha)(t - \xi_n^*).$$

Since S_R^n and S_D^n are non-decreasing processes,

$$\begin{aligned} & S_R^n(T_R^n(t)) - S_R^n(T_R^n(\xi_n^*-)) \\ & \geq S_R^n(T_R^n(\xi_n^*-)) + \alpha(t - \xi_n^*) - S_R^n(T_R^n(\xi_n^*-)) \\ & = \sqrt{n} \left[\tilde{S}_R^n(T_R^n(\xi_n^*-)) + \alpha(t - \xi_n^*) - \tilde{S}_R^n(T_R^n(\xi_n^*-)) + \alpha\sqrt{n}(t - \xi_n^*) \right], \end{aligned}$$

and

$$\begin{aligned} & S_D^n(T_D^n(t)) - S_D^n(T_D^n(\xi_n^*-)) \\ & \leq S_D^n(T_D^n(\xi_n^*-)) + (1 - \alpha)(t - \xi_n^*) - S_D^n(T_D^n(\xi_n^*-)) \\ & = \sqrt{n} \left[\tilde{S}_D^n(T_D^n(\xi_n^*-)) + (1 - \alpha)(t - \xi_n^*) - \tilde{S}_D^n(T_D^n(\xi_n^*-)) + (1 - \alpha)\sqrt{n}(t - \xi_n^*) \right]. \end{aligned}$$

The definition of ξ_n^* and substitution of the above upper bounds into (3.5) establish

$$\left| \frac{w_R}{\alpha} \tilde{Q}_R^n(t) - \frac{w_D}{1 - \alpha} \tilde{Q}_D^n(t) \right| \leq \frac{\epsilon}{2} + \tilde{U}_I^n(t, \xi_n^*, T_R^n(\xi_n^*), T_D^n(\xi_n^*)) + \frac{w_D}{1 - \alpha} \frac{1}{\sqrt{n}} R^n \left(\int_{\xi_n^*}^t \gamma Q_D^n(s) ds \right).$$

When $(w_R/\alpha)\tilde{Q}_R^n(\xi_n^*) \leq (w_D/(1 - \alpha))\tilde{Q}_D^n(\xi_n^*)$, a similar argument shows

$$\left| \frac{w_D}{1 - \alpha} \tilde{Q}_D^n(t) - \frac{w_R}{\alpha} \tilde{Q}_R^n(t) \right| \leq \frac{\epsilon}{2} + \tilde{U}_{II}^n(t, \xi_n^*, T_R^n(\xi_n^*), T_D^n(\xi_n^*)) - \frac{w_D}{1 - \alpha} \frac{1}{\sqrt{n}} R^n \left(\int_{\xi_n^*}^t \gamma Q_D^n(s) ds \right).$$

Also noting the process R^n is non-negative, we conclude

$$\begin{aligned} & \left| \frac{w_R}{\alpha} \tilde{Q}_R^n(t) - \frac{w_D}{1 - \alpha} \tilde{Q}_D^n(t) \right| \\ & \leq \frac{\epsilon}{2} + \max \left\{ \tilde{U}_I^n(t, \xi_n^*, T_R^n(\xi_n^*), T_D^n(\xi_n^*)), \tilde{U}_{II}^n(t, \xi_n^*, T_R^n(\xi_n^*), T_D^n(\xi_n^*)) \right\} \\ & \quad + \frac{w_D}{1 - \alpha} \frac{1}{\sqrt{n}} R^n \left(\int_0^T \gamma Q_D^n(s) ds \right). \end{aligned}$$

Therefore, the left-hand side of (3.4) can be bounded as follows

$$\begin{aligned} & P \left(\sup_{0 \leq t \leq T} \left| \frac{w_R}{\alpha} \tilde{Q}_R^n(t) - \frac{w_D}{1 - \alpha} \tilde{Q}_D^n(t) \right| > \epsilon \right) \\ & \leq P \left(\sup_{0 \leq s \leq t \leq T} \sup_{0 \leq u, v \leq s} \max \left\{ \tilde{U}_I^n(t, s, u, v), \tilde{U}_{II}^n(t, s, u, v) \right\} + \frac{w_D}{1 - \alpha} \frac{1}{\sqrt{n}} R^n \left(\int_0^T \gamma Q_D^n(s) ds \right) > \frac{\epsilon}{2} \right). \end{aligned} \tag{3.6}$$

Convergence of the right-hand-side of (3.6) to zero

Let η be arbitrarily small. Observe that

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

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

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

for all $n \geq n_0$.

The processes \tilde{A}^n , \tilde{S}_R^n , and \tilde{S}_D^n all weakly converge to Brownian motions by the functional central limit theorem. The heavy traffic assumption (2.8) implies that for any $t > s$, as $n \rightarrow \infty$,

$$\begin{aligned} \left(\frac{w_D}{1-\alpha} (\mu^n - \lambda^n) - \mu^n \left(w_R + \frac{\alpha}{1-\alpha} w_D \right) \right) \sqrt{n}(t-s) &\rightarrow -\infty \\ \left(\frac{w_R}{\alpha} (\mu^n - \lambda^n) - \mu^n \left(\frac{1-\alpha}{\alpha} w_R + w_D \right) \right) \sqrt{n}(t-s) &\rightarrow -\infty. \end{aligned}$$

Therefore, an argument analogous to the proof of Theorem 3.2 in Reiman [6] shows that there exists m_0 such that for all $n > m_0$

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

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

$$\begin{aligned} &P \left(\sup_{0 \leq s \leq t \leq T} \sup_{0 \leq u, v \leq s} \max \left\{ \tilde{U}_I^n(t, s, u, v), \tilde{U}_{II}^n(t, s, u, v) \right\} + \frac{w_D}{1-\alpha} \frac{1}{\sqrt{n}} R^n \left(\int_0^T \gamma Q_D^n(s) ds \right) > \frac{\epsilon}{2} \right) \\ &\leq P \left(\sup_{0 \leq s \leq t \leq T} \sup_{0 \leq u, v \leq s} \max \left\{ \tilde{U}_I^n(t, s, u, v), \tilde{U}_{II}^n(t, s, u, v) \right\} + M > \frac{\epsilon}{2} \right) \\ &\quad + P \left(\frac{w_D}{1-\alpha} \frac{1}{\sqrt{n}} R^n \left(\int_0^T \gamma Q_D^n(s) ds \right) > \frac{\epsilon}{2} \right) \\ &< \frac{\eta}{2} + \frac{\eta}{2} = \eta. \end{aligned}$$

3.2.2 Proof of Theorem 2

Define

$$\begin{aligned}\tilde{X}^n(t) &\equiv \tilde{A}^n(t) - \tilde{S}_R^n(\mu^n T_R^n(t)) - \tilde{S}_D^n(\mu^n T_D^n(t)) - \tilde{R}^n(\bar{\tau}^n(t)) + \sqrt{nt}(\lambda^n - \mu^n) \\ \tilde{\varepsilon}^n(t) &\equiv \gamma \int_0^t \left(\frac{(1-\alpha)w_I}{\alpha w_O + (1-\alpha)w_I} \tilde{Q}^n(s) - \tilde{Q}_D^n(s) \right) ds.\end{aligned}$$

Then, for all $t \geq 0$,

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

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

$$\left(\tilde{Q}^n, \tilde{I}^n \right) \equiv (\phi^\kappa, \psi^\kappa) \left(\tilde{X}^n + \tilde{\varepsilon}^n \right). \quad (3.7)$$

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

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

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

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

as $n \rightarrow \infty$. Proposition 4 part (iii) in Ward and Kumar [8] establishes the mapping $(\phi^\kappa, \psi^\kappa)$ is continuous. Therefore, by the continuous mapping theorem

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

as $n \rightarrow \infty$. The representation (3.7) then establishes

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

as $n \rightarrow \infty$.

The sequence $\{(T_D^n, T_R^n)\}$ is tight in D because $|T_\Delta^n(t) - T_\Delta^n(s)| \leq |t - s|$ for $\Delta \in \{R, D\}$. Consider any subsequence $\{n_k\}$ on which

$$(T_D^{n_k}, T_R^{n_k}) \Rightarrow (T_D, T_R)$$

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

$$T_D + T_R = e.$$

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 (2.8)

$$\begin{aligned} & \tilde{A}^{n_k}(t) - \tilde{S}_R^{n_k}(\mu^{n_k} T_R^{n_k}(t)) - \tilde{S}_D^{n_k}(\mu^{n_k} T_D^{n_k}(t)) + \sqrt{n_k} t (\lambda^{n_k} - \mu^{n_k}) \\ & \Rightarrow \sqrt{\text{var}(u_1)} B_1 - \sqrt{\text{var}(v_1^R)} B_2 \circ T_R - \sqrt{\text{var}(v_1^D)} B_3 \circ T_D + \theta e, \end{aligned}$$

as $n_k \rightarrow \infty$. By the same argument directly following (3.6) in the proof of Theorem 1,

$$\tilde{R}^{n_k} \circ \bar{\tau}^{n_k}(t) \Rightarrow 0,$$

as $n_k \rightarrow \infty$. Therefore,

$$\tilde{X}^{n_k} \Rightarrow \sqrt{\text{var}(u_1)} B_1 - \sqrt{\text{var}(v_1^R)} B_2 \circ T_R - \sqrt{\text{var}(v_1^D)} B_3 \circ T_D + \theta e,$$

as $n_k \rightarrow \infty$. Since $T_R + T_D = e$ and $\text{var}(v_1^R) = \text{var}(v_1^D)$ by assumption, it follows that $\text{var}(u_1) B_1 - \text{var}(v_1^R) B_2 \circ T_R - \text{var}(v_1^D) B_3 \circ T_D$ has the same distribution as σW . Since the subsequence $\{n_k\}$ was arbitrary, we conclude

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

as $n \rightarrow \infty$. □

3.2.3 Proof of Theorem 3

We establish

$$\tilde{W}_D^n \Rightarrow \frac{(1-\alpha)w_R}{(1-\alpha)w_R + \alpha w_D} \frac{Z}{\mu} \tag{3.8}$$

as $n \rightarrow \infty$. Showing

$$\tilde{W}_R^n \Rightarrow \frac{\alpha w_D}{(1-\alpha)w_R + \alpha w_D} \frac{Z}{\mu}$$

as $n \rightarrow \infty$ follows an argument similar to Theorem 5.3 in Reiman [5], and so is omitted. Since the delayed service queue receives at least $(1-\alpha)$ proportion

of the server's efforts when the queue is non-empty, $(1 - \alpha)^{-1}W_D^n(t)$ exceeds the amount of time required to finish serving all customers in the delayed service queue that will eventually receive service. Therefore, at time $t > 0$, the number of customers in the delayed service queue that will eventually renege is less than or equal to

$$\mathcal{R}^n(t) \equiv R \left(\int_0^{t+(1-\alpha)^{-1}W_D^n(t)} \gamma Q_D^n(s) ds \right) - R \left(\int_0^t \gamma Q_D^n(s) ds \right).$$

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

$$L_D^n(t) \equiv \sum_{j=S_D^n(T_D^n(t))+2}^{S_D^n(T_D^n(t))+Q_D^n(t)-\mathcal{R}^n(t)} \frac{v_j^D}{n\mu^n} \leq W_D^n(t).$$

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

$$U_D^n(t) \equiv \sum_{j=S_D^n(T_D^n(t))+1}^{S_D^n(T_D^n(t))+Q_D^n(t)} \frac{v_j^D}{n\mu^n} \geq W_D^n(t).$$

We conclude

$$0 \leq \sqrt{n}W_D^n(t) - \sqrt{n}L_D^n(t) \leq \sqrt{n}U_D^n(t) - \sqrt{n}L_D^n(t). \quad (3.9)$$

Define

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

Observe that

$$\begin{aligned} & \sqrt{n}U_D^n(t) - \sqrt{n}L_D^n(t) \\ &= \frac{1}{\mu^n} \frac{1}{\sqrt{n}} v_{S_D^n(T_D^n(t))+1}^D + \frac{1}{\mu^n} \frac{1}{\sqrt{n}} \mathcal{R}^n(t) \\ & \quad + \frac{1}{\mu^n} \left(\tilde{V}_D^n \left(\frac{S_D^n(T_D^n(t))}{n} + \frac{Q_D^n(t)}{n} \right) - \tilde{V}_D^n \left(\frac{S_D^n(T_D^n(t))}{n} + \frac{Q_D^n(t)}{n} - \frac{\mathcal{R}^n(t)}{n} \right) \right) \end{aligned} \quad (3.10)$$

and

$$\begin{aligned}
& \sqrt{n}L_D^n(t) \tag{3.11} \\
&= \frac{1}{\mu^n} \tilde{Q}_D^n(t) - \frac{1}{\mu^n} \frac{1}{\sqrt{n}} - \frac{1}{\mu^n} \frac{1}{\sqrt{n}} \mathcal{R}^n(t) \\
& \quad + \frac{1}{\mu^n} \left(\tilde{V}_D^n \left(\frac{S_D^n(T_D^n(t))}{n} + \frac{Q_D^n(t)}{n} - \frac{\mathcal{R}^n(t)}{n} \right) - \tilde{V}_D^n \left(\frac{S_D^n(T_D^n(t))}{n} + \frac{1}{n} \right) \right).
\end{aligned}$$

We will first show that $\sqrt{n}U_D^n - \sqrt{n}L_D^n \Rightarrow 0$ as $n \rightarrow \infty$, and then show

$$\sqrt{n}L_D^n \Rightarrow \frac{(1-\alpha)w_R}{(1-\alpha)w_R + \alpha w_D} \frac{Z}{\mu} \tag{3.12}$$

as $n \rightarrow \infty$. The inequality (3.9) and the converging together lemma then establish (3.8) and complete the proof.

Since

$$\frac{1}{\sqrt{n}} \mathcal{R}^n(t) = \tilde{R}^n \left(\bar{\tau}^n \left(t + \frac{W_D^n(t)}{1-\alpha} \right) \right) - \tilde{R}^n(\bar{\tau}^n(t)) + \int_t^{t+(1-\alpha)^{-1}W_D^n(t)} \gamma \tilde{Q}_D^n(s) ds,$$

and Lemma 1 establishes $\bar{\tau}^n \rightarrow 0$ and $W_D^n \rightarrow 0$ a.s., u.o.c., it follows from the functional central limit theorem, continuous mapping theorem, and the weak convergence of \tilde{Q}_D^n in (3.3) that

$$\frac{1}{\sqrt{n}} \mathcal{R}^n \Rightarrow 0 \tag{3.13}$$

as $n \rightarrow \infty$. By Lemma 3 in Iglehart and Whitt [2],

$$\sup_{1 \leq k \leq S_D^n(n\mu^n)+1} \frac{v_k}{\sqrt{n}} \rightarrow 0$$

in probability, as $n \rightarrow \infty$. The sequence $\{T_D^n\}$ is tight in D because $|T_D^n(t) - T_D^n(s)| \leq |t - s|$. On any subsequence $\{n_k\}$ on which

$$T_D^{n_k} \Rightarrow T_D$$

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

$$\frac{S_D^{n_k} \circ T_D^{n_k}}{n_k} \Rightarrow \mu T_D$$

as $n_k \rightarrow \infty$. Because by Donsker's theorem \tilde{V}_D^n weakly converges to a continuous limit process, and by the convergences in (3.13) and Lemma 1, $n_k^{-1}\mathcal{R}^{n_k} \Rightarrow 0$ as $n_k \rightarrow \infty$, $n_k^{-1}Q_D^{n_k} \rightarrow 0$ a.s., u.o.c. as $n_k \rightarrow \infty$, it follows that

$$\tilde{V}_D^{n_k} \left(\frac{S_D^{n_k}(T_D^{n_k}(\cdot))}{n_k} + \frac{Q_D^{n_k}(\cdot)}{n_k} \right) - \tilde{V}_D^{n_k} \left(\frac{S_D^{n_k}(T_D^{n_k}(\cdot))}{n_k} + \frac{Q_D^{n_k}(\cdot)}{n_k} - \frac{\mathcal{R}^{n_k}(\cdot)}{n_k} \right) \Rightarrow 0$$

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

$$\tilde{V}_D^n \left(\frac{S_D^n(T_D^n(\cdot))}{n} + \frac{Q_D^n(\cdot)}{n} \right) - \tilde{V}_D^n \left(\frac{S_D^n(T_D^n(\cdot))}{n} + \frac{Q_D^n(\cdot)}{n} - \frac{\mathcal{R}^n(\cdot)}{n} \right) \Rightarrow 0$$

as $n \rightarrow \infty$. We conclude from (3.10) that

$$\sqrt{n}U_D^n - \sqrt{n}L_D^n \Rightarrow 0$$

as $n \rightarrow \infty$.

We now establish (3.12). An argument similar to that in the above paragraph shows

$$\tilde{V}_D^n \left(\frac{S_D^n(T_D^n(\cdot))}{n} + \frac{Q_D^n(\cdot)}{n} - \frac{\mathcal{R}^n(\cdot)}{n} \right) - \tilde{V}_D^n \left(\frac{S_D^n(T_D^n(\cdot))}{n} + \frac{1}{n} \right) \Rightarrow 0$$

as $n \rightarrow \infty$. Hence, the representation of $\sqrt{n}L_D^n$ in (3.11), Theorems 1 and 2 (specifically, the resulting convergence in (3.3)), the convergence in (3.13), and the continuous mapping theorem establish the weak convergence in (3.12) required to complete the proof. \square

4 Revenue Optimization

5 A Numerical Study

[NOTE: WE SHOULD ALSO BE ABLE TO APPROXIMATE THE PROBABILITY A CUSTOMER IN THE OFFLINE QUEUE ABANDONS.]

References

- [1] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, Inc., New York, 1999. Second Edition.

- [2] D. L. Iglehart and W. Whitt. Multiple channels queues in heavy traffic i. *Adv. in Applied Probability*, 2:150–177, 1970.
- [3] L. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve. An explicit formula for double reflected processes in $[0,a]$, 2005. Submitted.
- [4] J. Reed and A. R. Ward. Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic, 2006. Working Paper.
- [5] M. I. Reiman. The heavy traffic diffusion approximation for sojourn times in Jackson networks. In R. L. Disney and T. J. Ott, editors, *Applied Probability-Computer Science, The Interface*. Birkhauser, Boston, 1982.
- [6] M. I. Reiman. Some diffusion approximations with state space collapse. In F. Bacceli and G. Fayolle, editors, *Modelling and Performance Evaluation Methodology*, pages 209–240. Springer-Verlag, 1984.
- [7] A. V. Skorokhod. Stochastic equations for diffusions in a bounded region. *Theor. of Prob. and Its Appl.*, 6:264–274, 1961.
- [8] A. R. Ward and S. Kumar. Asymptotically optimal control of a queue with impatient customers, 2007. Forthcoming in *Mathematics of Operations Research Research*.
- [9] W. Whitt. *Stochastic-Process Limits*. Springer, New York, 2002.

A Proofs of Lemmas 1-2

Proof of Lemma 1

Define

$$\bar{X}^n(t) \equiv \bar{A}^n(t) - \bar{S}_R^n(I_R^n(t)) - \bar{S}_D^n(I_D^n(t)) - \bar{R}^n(\bar{\tau}^n(t)) + (\lambda^n - \mu^n)t.$$

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 $\int_0^\infty \bar{Q}^n(t) d(\mu^n I^n(t)) = 0$, the process $(\bar{Q}^n, \mu^n I^n)$ can be represented in terms of the conventional two-sided regulator mapping as follows

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

Since $\bar{\tau}^n$ is a non-decreasing process, Lemma 5.1 in Kruk et al [3] establishes

$$\phi(\bar{X}^n - \bar{\tau}^n) \leq \phi(\bar{X}^n).$$

The functional strong law of large numbers and the heavy traffic assumption (2.8) establish

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

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

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

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

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

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

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

as $n \rightarrow \infty$, and so

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

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

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

as $n \rightarrow \infty$. The condition (2.6) then implies

$$T_R^n + T_D^n \rightarrow e \text{ a.s., u.o.c.,}$$

as $n \rightarrow \infty$.

It remains to show

$$W_D^n \rightarrow 0 \text{ a.s., u.o.c.,}$$

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

$$U_D^n(t) \equiv \sum_{j=S_D^n(T_D^n(t))+1}^{S_D^n(T_D^n(t))+Q_D^n(t)} \frac{v_j^D}{n\mu^n}$$

defined as in the proof of Theorem 3,

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

Define

$$\bar{V}_D^n(t) \equiv \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} (v_i^D - 1),$$

and observe that

$$U_D^n(t) = \frac{1}{\mu^n} \left(\bar{V}_D^n \left(\frac{1}{n} S_D^n(T_D^n(t)) + \bar{Q}_D^n(t) \right) - \bar{V}_D^n \left(\frac{1}{n} S_D^n(T_D^n(t)) \right) \right) + \frac{1}{\mu^n} \bar{Q}_D^n(t).$$

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

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

as $n \rightarrow \infty$. Therefore, because $\bar{U}_D^n \rightarrow 0$ a.s., u.o.c. as $n \rightarrow \infty$, we conclude

$$W_D^n \rightarrow 0 \text{ a.s., u.o.c.,}$$

as $n \rightarrow \infty$. □

Proof of Lemma 2

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

$$\begin{aligned} \tilde{\chi}^n &\equiv \tilde{A}^n(t) - \tilde{S}_R^n(T_R^n(t)) - \tilde{S}_D^n(T_D^n(t)) + \sqrt{nt}(\lambda^n - \mu^n) \\ \tilde{\mathcal{R}}^n(t) &\equiv \frac{1}{\sqrt{n}} R \left(\int_0^t \gamma Q_D^n(s) ds \right). \end{aligned}$$

Then,

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

Since \tilde{I}^n is non-decreasing, $\tilde{I}^n(0) = 0$, and the condition (2.7) implies $\int_0^\infty \tilde{Q}^n(t) d\tilde{I}^n(t) = 0$, the process $(\tilde{Q}^n, \mu^n \tilde{I}^n)$ can be represented in terms of the conventional two-sided regulator mapping as follows

$$(\tilde{Q}^n, \mu^n \tilde{I}^n) = (\phi, \psi) (\tilde{\chi}^n - \tilde{\mathcal{R}}^n). \quad (\text{A.1})$$

Since $\tilde{\mathcal{R}}^n$ is a non-decreasing process, Lemma 5.1 in Kruk et al [3] establishes that

$$\phi\left(\tilde{\chi}^n - \tilde{\mathcal{R}}^n\right)(t) \leq \phi\left(\tilde{\chi}^n\right)(t) \text{ for all } t \geq 0. \quad (\text{A.2})$$

The functional central limit theorem, continuous mapping theorem, and heavy traffic assumption (2.8) establish

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

as $n \rightarrow \infty$. 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 (A.1) and the upper bound (A.2) that

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

□