Join the Shortest Queue with Many Servers.
The Heavy Traffic Asymptotics

Patrick Eschenfeldt
MIT Operations Research Center

David Gamarnik
MIT Sloan School of Management and Operations Research Center

Abstract

We consider queueing systems with \( n \) parallel queues under a Join the Shortest Queue (JSQ) policy in the Halfin-Whitt heavy traffic regime. Because queues with at least two customers form only when all queues have at least one customer and we expect the number of waiting customers to be of the order \( O(\sqrt{n}) \), we restrict our attention to a truncated system that rejects arrivals creating queues longer than two. We provide simulation results supporting this intuition. We use the martingale method to prove that a scaled process counting the number of idle servers and queues of length 2 weakly converges to a two-dimensional reflected Ornstein-Uhlenbeck process. This limiting system is comparable to that of the traditional Halfin-Whitt model, but there are key differences in the queueing behavior of the JSQ model. In particular, it is possible for the system to have both idle servers and waiting customers at the same time.

1 Introduction.

In this paper we consider queueing systems with many parallel servers under a heavy-traffic regime where the workload scales with the number of servers. Such systems are well understood when a global queue is maintained \cite{10}, but in many practical situations it may be advantageous to instead maintain parallel queues. Even if a global queue is itself not problematic, it may be necessary to keep queued customers close to the server who will eventually serve them. Consider for example an airport setting with arriving passengers who need to have their passports checked with one of a large number of passport controllers. In this situation having only a global queue can lead to significant walk times between the front of the queue and the server, leaving servers idle while they wait for their next customer. This idle time can be avoided by routing customers to individual queues for each server before earlier customers finish service.

At the same time, a parallel scheme will necessarily allow servers to idle if their own queue is empty, even if customers are waiting in another queue, thus sacrificing some efficiency. We will study a particular example of a parallel system and demonstrate that under certain (Halfin-Whitt) heavy traffic conditions for this system, the number of idle servers is on the order \( O(\sqrt{n}) \), where \( n \) is the number of servers.

Specifically, this result will be for a parallel queueing system in which each arriving customer is immediately routed to the queue containing the smallest number of customers, namely the Join the Shortest Queue (JSQ) policy. We consider the system in heavy traffic by allowing the arrival rate \( \lambda_n \) to depend on \( n \). In particular we consider the Halfin-Whitt regime \cite{10}, defined by letting the quantity \((1 - \lambda_n)\sqrt{n}\) having a non-degenerate limit, which we will denote \( \beta > 0 \). Note that JSQ is a logical first step for understanding the tradeoffs involved in maintaining parallel...
queues, because Winston [20] proved that among policies immediately assigning customers to one of \( n < \infty \) parallel queues, JSQ is optimal in the case of Poisson arrivals and exponential service times. That is, it maximizes, with respect to stochastic order, the number of customers served in a given time interval. Weber [18] extended this result to the more general class of service times with non-decreasing hazard rate, with no assumptions on the arrival process. We will consider Poisson arrivals and exponential service times, and denote this system \( M/M/n-JSQ \), distinguishing it from the traditional \( M/M/n \) system which maintains a global queue.

Our main result describes the behavior of processes counting the number of idle servers and queues with a customer waiting to enter service. These processes will be analyzed in a truncated variant of the \( M/M/n-JSQ \) system in which no queue may have length longer than 2, and thus each server can have at most one customer waiting to enter service. We use this simplified system because longer queues form only when all queues are length 2. We conjecture that in steady state the number of queues with length 2 will be on the order \( O(\sqrt{n}) \), so with high probability longer queues will not exist. While we are not able to verify this conjecture, our intuition is supported by simulation results. In particular, we simulate the system which starts with no queues of length 3 or longer and a small number of queues with length 2. In these simulations, longer queues do not appear. See Figure 1 for a simulated sample path.

![Figure 1: A sample path of the \( M/M/n-JSQ \) system scaled by \( \sqrt{n} \). Simulated with \( n = 10^5 \), \( \beta = 2.0 \), started with \( 3\beta\sqrt{n} \) idle servers, \( 3\beta\sqrt{n} \) queues of length two and the remaining servers with a single customer in service. Plot shows idle servers \((-X_1)\) and queues of length two \((X_2)\) as a function of time.](image)

Thus we believe that the truncated system, in which arrivals are rejected if they would create a queue of length 3, will be an accurate representation of the number of idle servers and waiting customers in the original system. Therefore in this paper we consider a sequence of truncated \( M/M/n-JSQ \) systems in the Halfin-Whitt regime and prove that this sequence converges weakly to a diffusion process. This diffusion process, which is a multidimensional reflected Ornstein-
Uhlenbeck process, will be defined in terms of a stochastic integral equation which we prove has a unique solution. This existence and uniqueness result is stated in Theorem 1 and the weak convergence result is stated in Theorem 2, which is our main result.

A feature of interest in this queueing system is the waiting time experienced by arriving customers. We conjecture that in steady state the waiting time $W$ of a typical customer behaves as follows: $P(W = 0) = 1 - O(1/\sqrt{n})$, and conditioned on $W > 0$ (which thus occurs with probability $O(1/\sqrt{n})$, it is exponential with parameter 1. Our heuristic reasoning is as follows. Because customers do not jockey between queues, when the customer joins the server with another customer in service, the waiting time of this arriving customer is at least of constant order. Notice, on the other hand, that an arriving customer immediately enters service if there are any idle servers in the system. In this case the incurred waiting time is zero. Thus the expected waiting time can be characterized by the fraction of arrivals which occur when there are no idle servers. In the limiting system this will be the local time at zero of a reflected diffusion process, which is zero in the limit. Based on that we conjecture that in steady state the probability of a customer having to wait will be of the order $O(1/\sqrt{n})$, in which case the conditioned waiting time equals unity, because it corresponds to time to finish servicing the preceding customer (service times are exponential and normalized to unity). Simulation results support this intuition. In particular, the fraction of arrivals which find no idle server in simulated sample paths multiplied by $\sqrt{n}$ tends to a constant as $n$ increases. This conjecture if true would imply that the expected waiting time of the truncated JSQ system is on the same order as that of the $M/M/n$ system, so the inefficiency of allowing servers to idle while customers wait does not result in an order of magnitude change in the expected waiting time.

1.1 Related literature.

The JSQ model was initially studied in the special case of 2 queues by Haight [8]. Kingman [12] proved stability results along with considering the stationary distribution of the system, and Flatto and McKean [5] also examine the stationary distribution. Further work on the $n = 2$ case includes bounds on the distribution of the number of people in the system by Halfin [9].

Foschini and Salz [6] consider diffusion limits for the heavy traffic case of the $M/M/2$-JSQ system, first proving that the queue-length processes for the two queues are identical in the limit and then deriving the limiting distribution. The limiting behavior of the waiting time is the same as the standard $M/M/2$ system in heavy traffic. Their results extend to the case of $k$ parallel queues, but they do not consider the case where the number of queues grows as the traffic intensity increases. Zhang and Wang [11] and Zhang and Hsu [21] look at a similar problem but drop the assumption of Poisson arrivals and exponential service times, deriving functional central limit theorems for the heavy traffic JSQ system with $s$ servers.

In contrast to much early work on JSQ systems which consider only a fixed number of servers, we are interested in the asymptotics as the number of servers $n$ increases. Our first observation is that for fixed $\lambda < 1$ as $n$ increases the probability of any customer arriving to find all servers busy will decrease to zero. In this case the JSQ nature of the system becomes irrelevant as customers will be assigned to an idle server immediately upon arrival. In particular we see that the limiting behavior of the system will essentially be that of the well known $M/M/\infty$ system, and thus it is of interest to consider this model in heavy traffic with $\lambda$ approaching unity. There has been some work on models similar to ours, most notably Tezcan [16], who considers a variant of the JSQ system with multiple pools of servers who each have their own queue. He uses a state-space collapse argument based on a framework of Dai and Tezcan [2] to prove diffusion limits under the Halfin-Whitt heavy traffic regime. In this case that regime has the number of servers and traffic intensity increasing together in the limit, but the number of pools of servers is fixed so the number of queues is also fixed. Therefore our model is similar to Tezcan’s but is not a special case of it. The state-space collapse argument implies that in the limit the system can
be fully described by the total number of people in the system (rather than the queue lengths in the individual pools) and the diffusion limit of that process is very similar to the original Halfin and Whitt result [10].

Another branch of analysis of JSQ-like queueing systems has focused on the “supermarket model” in which arriving customers join the shortest queue from among \(d\) randomly selected queues rather than from the entire system. It was proved independently by Mitzenmacher [13] and Vvedenskaya, Dobrushin, and Karpelevich [17] that this system achieves an exponential improvement in expected waiting time over a system with \(n\) independent \(M/M/1\) queues. Versions of this system where \(d\) depends on \(n\) are particularly closely related to our JSQ model, which essentially sets \(d = n\). Brightwell and Luczak [1] give a set of \(d\) and \(\lambda\) values depending on \(n\) for which they prove the steady-state system is usually in a particular state with most queues having the same (known) length. Their conditions require \((1 − \lambda)^{−1} > d\), which excludes the \(d = n, (1 − \lambda)\sqrt{n} \to \beta\) case considered in this paper. Dieker and Suk [4] prove fluid and diffusion limits for queue length processes when \(d\) increases to infinity at a rate slower than \(n\) and with fixed \(\lambda < 1\).

The remainder of the paper is laid out as follows: Section 2 will define the model and state our main result. Section 3 will prove Theorem 1, verifying that the integral representation of the limiting system is well defined. This result will also be the key to proving convergence via a continuous mapping theorem (CMT) argument. Section 4 will construct a representation of the system as a combination of martingales and reflecting processes. In Section 5 we will establish the convergence properties of these martingales, and then apply the CMT to translate the convergence of martingales to convergence of the scaled queue length processes. This section will conclude our proof of Theorem 2. We will conclude in Section 6 with a brief discussion of the implications of Theorem 2 and possible extensions.

We use \(\Rightarrow\) to denote weak convergence, \(\mathbb{1}\{A\}\) to denote the indicator function for the event \(A\), \((x)^+ = \max(x, 0)\). We let \(\mathbb{R}_+ = \mathbb{R}_+ \cup \{\infty\}\) represent the extended positive real line. Most processes in this paper will live in the space \(D = D([0, \infty), \mathbb{R})\) of right continuous functions with left limits mapping \([0, \infty)\) into \(\mathbb{R}\). We also consider \(D^k = D([0, \infty), \mathbb{R}^k)\) for \(k \geq 2\), which we will treat as the product space \(D \times D \times \cdots \times D\) (see, e.g., [19] §3.3). We will denote the uniform norm 
\[
||x||_t = \sup_{0 \leq s \leq t} |x(s)|
\]
for \(x \in D\) and the max norm 
\[
||(x_1, \ldots, x_k)||_t = \max_{1 \leq i \leq k} ||x_i||_t
\]
for \(x \in D^k\). Similarly we will use the max norm 
\[
|b| = \max_{1 \leq i \leq k} |b_i|
\]
for \(b \in \mathbb{R}^k\).

## 2 The model, simulations, and the main result.

We consider a queueing system with \(n\) servers where each server maintains a unique queue, with service proceeding according to a first-in-first-out discipline. Service time is exponentially distributed at each server, with the rate fixed at 1. Arrivals occur in a single stream, as a Poisson process with rate \(\lambda_n n\), where \(0 < \lambda_n < 1\) and 
\[
\lim_{n \to \infty} \sqrt{n}(1 − \lambda_n) = \beta
\]
(2.1)
for fixed $\beta > 0$. Upon arrival, each customer is routed to the server with the shortest queue. In
the event of a tie, one of the options is selected uniformly at random.

The state of the system will be represented primarily via the process $Q^n(t) = (Q^n_1(t), Q^n_2(t), \ldots)$, with $Q^n_k(t)$ representing the number of queues with at least $k$ customers (including any customer in service) at time $t \geq 0$. We note that for the system as described we have

$$n \geq Q^n_1(t) \geq Q^n_2(t) \geq \cdots \geq 0 \quad \forall t \geq 0,$$

and that we can recover the number of queues with exactly $k$ customers in service via the quantity $Q^n_k(t) - Q^n_{k+1}(t)$, including the number of idle servers $n - Q^n_1(t)$.

To state our weak convergence results, we also introduce a scaled version $X^n(t)$ of this process defined as

$$X^n_1(t) = \frac{Q^n_1(t) - n}{\sqrt{n}} \quad \text{and} \quad X^n_k(t) = \frac{Q^n_k(t)}{\sqrt{n}}.$$

The $k = 1$ case is treated differently because the number of queues with length 1 behaves
differently than the number of queues of all larger lengths, as we are about to see.

We expect, both through an intuitive consideration of the system and through simulations,
that outside of the effect of the starting state the relevant behavior will be captured by only
considering queues of length at most two. Specifically, we expect the number of idle servers and
number of queues with length two is order $\Theta(\sqrt{n})$ while queues with length three or more will
be present only if they are present in the initial condition. This is because such long queues only
form when all $n$ queues have length 2, but the number of queues with length 2 is $\Theta(\sqrt{n})$. See
Figure 2 to see a sample path of an untruncated system which starts with order $\Theta(\sqrt{n})$ queues
of length 4. Note that the number of queues length 3 and length 4 decrease monotonically.

![Figure 2: A simulated sample path of an untruncated M/M/n-JSQ system, showing the scaled number of idle servers (a) and queues of length at least two (X_2), three (X_3), four (X_4), and five (X_5). Simulated with n = 10^5, \beta = 2.0.](image)

Considering all possible queue lengths $k \geq 3$ complicates notation and technical aspects of
proofs, so we will exclusively consider a modification of the system in which long queues are
excluded by design. This will allow us to demonstrate weak convergence to a particular diffusion
process.

Formally, any arrivals while all servers have a queue of length two are rejected. In this
modified system we need only consider $Q^n_1$ and $Q^n_2$ (and their scaled counterparts). We note
that \((Q^n_1, Q^n_2)\) is a member of the function space \(D^2\). Before we state our primary result, we will state the theorem that demonstrates the existence and uniqueness of the diffusion limit we will prove.

Our diffusion limit will be the solution to a system of integral equations, so we first prove that the system has a unique solution. Furthermore, we prove that the system defines a continuous map from \(\mathbb{R}_+ \times \mathbb{R}^2 \times D^2\) to \(D^2 \times D^2\) with respect to appropriate topologies. This continuity, along with the further fact that the function maps continuous functions to continuous functions, allows us to use the CMT to prove weak convergence once we show the weak convergence of the arguments. Because our limiting system has continuous sample paths, we equip \(D\) with the topology of uniform convergence over bounded intervals.

**Theorem 1.** Given \(B \in \mathbb{R}_+, b \in \mathbb{R}^2, \text{ and } y \in D^2\), consider
\[
\begin{align*}
x_1(t) &= b_1 + y_1(t) + \int_0^t (-x_1(s) + x_2(s))ds - u_1(t), \quad (2.4) \\
x_2(t) &= b_2 + y_2(t) + \int_0^t (-x_2(s))ds + u_1(t) - u_2(t), \quad (2.5) \\
x_1(t) &\leq 0, \quad 0 \leq x_2(t) \leq B, \quad t \geq 0, \quad (2.6)
\end{align*}
\]

with \(u_1\) and \(u_2\) nondecreasing nonnegative functions in \(D\) such that
\[
\int_0^\infty \mathbb{1}\{x_1(t) < 0\}du_1(t) = 0,
\]
\[
\int_0^\infty \mathbb{1}\{x_2(t) < B\}du_2(t) = 0.
\]

Then \((2.4) - (2.6)\) has a unique solution \((x, u) \in D^2 \times D^2\) so that there is a well defined function \((f, g) : \mathbb{R}_+ \times \mathbb{R}^2 \times D^2 \to D^2 \times D^2\) mapping \((B, b, y)\) into \(x = f(B, b, y)\) and \(u = g(B, b, y)\). Furthermore, the function \((f, g)\) is continuous on \(\mathbb{R}_+ \times \mathbb{R}^2 \times D^2\). Finally, if \(y\) is continuous, then so are \(x\) and \(u\).

We will prove this theorem in Section 3. One implication of Theorem 1 is that the limiting system we find in our main result is well defined because it is an application of the function \((f, g)\) with specific arguments \((B, b, y)\). Note in particular that we will have \(B = \infty\), which implies \(u_2 = 0\). Our main result is the following:

**Theorem 2.** In the sequence of truncated JSQ models described above, suppose that
\[
X^n_k(0) \Rightarrow X_k(0) \quad \text{in } \mathbb{R} \text{ as } n \to \infty, \quad k = 1, 2. \quad (2.7)
\]

Then
\[
X^n_k \Rightarrow X_k \quad \text{in } D \text{ as } n \to \infty, \quad k = 1, 2,
\]

where \(X_1 \leq 0\) and \(X_2 \geq 0\) are unique solutions in \(D\) for the stochastic integral equations
\[
\begin{align*}
X_1(t) &= X_1(0) + \sqrt{2}W(t) - \beta t + \int_0^t (-X_1(s) + X_2(s))ds - U_1(t), \quad (2.8) \\
X_2(t) &= X_2(0) + U_1(t) + \int_0^t -X_2(s)ds, \quad (2.9)
\end{align*}
\]

for \(t \geq 0\) where \(W\) is a standard Brownian motion and \(U_1\) is the unique nondecreasing nonnegative process in \(D\) satisfying
\[
\int_0^\infty \mathbb{1}\{X_1(t) < 0\}dU_1(t) = 0. \quad (2.10)
\]
We note that condition (2.7) does place significant but not unreasonable restrictions on the starting state of the finite systems $Q^n$. In particular, $Q^n(0) - n = O(\sqrt{n})$ so the number of customers initially in service must be sufficiently near $n$. Similarly, (2.7) requires $Q^n(0) = O(\sqrt{n})$.

3 Integral representation.

We will now prove Theorem 1, showing that the representation of the limiting system in Theorem 2 is a valid and unique representation. We will also show that it defines a continuous map from $\mathbb{R}_+ \times \mathbb{R}^2 \times D^2$ to $D^2 \times D^2$. The continuity of the map in the topology of uniform convergence over bounded intervals will allow us to use the continuous mapping theorem (CMT) to demonstrate the convergence $X^n_k \Rightarrow X^k$ once we write $X^n_k$ in the appropriate integral form.

Note that by using $\mathbb{R}_+$ in the domain of this map we allow the upper barrier $B$ for the function $x_2$ to take the value $\infty$, which corresponds to there being no upper barrier on the $\sqrt{n}$ scale.

3.1 The reflection map.

In several places we will make use of the well known one-dimensional reflection map for an upper barrier. Given upper barrier $\kappa \in \mathbb{R}_+$, we let $(\phi_\kappa, \psi_\kappa) : D \rightarrow D^2$ be the one-sided reflection map with upper barrier at $\kappa$ (see, e.g., [19] §5.2 and §13.5). In particular for $x \in D$ with $x(0) \leq \kappa$ we have $z = \psi_\kappa(y) \geq 0$, $z$ nondecreasing, $x = \phi_\kappa(y) = y - z \leq \kappa$, and

$$\int_0^\infty 1\{x < \kappa\}dz = 0.$$  

Recall that these functions can be defined explicitly by

$$\psi_\kappa(x)(t) = \sup_{0 \leq s \leq t} (x(s) - \kappa)^+$$ (3.1)

and

$$\phi_\kappa(x)(t) = x(t) - \psi_\kappa(x)(t).$$ (3.2)

We will also make use of a slight variant of the usual Lipschitz condition for these functions to allow for different values of $\kappa$. In particular, for $x, x' \in D, \kappa, \kappa' \in \mathbb{R}$, and $t \geq 0$ we have

$$||\psi_\kappa(x) - \psi_\kappa(x')||_t \leq ||x - x'||_t + |\kappa - \kappa'|,$$ (3.3)

$$||\phi_\kappa(x) - \phi_\kappa(x')||_t \leq 2 ||x - x'||_t + |\kappa - \kappa'|.$$ (3.4)

These follow straightforwardly from (3.1) and (3.2). Note that for $\kappa = \kappa'$ we recover the usual Lipschitz constants of $1$ for $\psi_\kappa$ and $2$ for $\phi_\kappa$.

We also define a trivial reflection map for $\kappa = \infty$ by letting $(\phi_\infty, \psi_\infty) = (e, 0)$ where $e$ is the identity map. That is, the reflection map leaves the argument unchanged and the regulator is identically zero. We prove the following:

**Lemma 1.** The function $(\phi, \psi) : \mathbb{R}_+ \times D \rightarrow D^2$ defined by (3.1)-(3.2) for finite $\kappa$ and by $(\phi_\infty, \psi_\infty) = (e, 0)$ for $\kappa = \infty$ is continuous with respect to the product topology when $\mathbb{R}_+$ is equipped with the order topology and $D$ is equipped with the topology of uniform convergence over bounded intervals.
Proof. By (3.3)-(3.4) the function is continuous at any finite $\kappa \in \mathbb{R}$. For $x \in D$ and $x^\kappa \in D$ such that $x^\kappa \to x$ as $\kappa \to \infty$,

$$
\lim_{\kappa \to \infty} ||\psi_\kappa(x^\kappa)||_t = \lim_{\kappa \to \infty} \sup_{0 \leq s \leq t} |\psi_\kappa(x^\kappa)|
= \lim_{\kappa \to \infty} \sup_{0 \leq s \leq t} (x^\kappa(s) - \kappa)^+
= \sup_{0 \leq s \leq t} \lim_{\kappa \to \infty} (x^\kappa(s) - \kappa)^+
= 0,
$$

where we have made use of the fact that $||x||_t < \infty$. Therefore $\psi_\kappa(x^\kappa) \to \psi_\infty(x)$ and by (3.2) we conclude $\phi_\kappa(x^\kappa) \to \phi_\infty(x)$. Thus the function is continuous at $\kappa = \infty$, completing the proof. \(\square\)

With these facts about the reflection map in hand, we will now prove a result similar to Theorem 1 for a related system:

**Lemma 2.** Given $B \in \mathbb{R}_+$, $b \in \mathbb{R}^2$ and $y \in D^2$, consider

$$w_1(t) = b_1 + y_1(t) + \int_0^t (-\phi_0(w_1(s)) + \phi_B(w_2(s))) \, ds, \quad (3.5)$$

$$w_2(t) = b_2 + y_2(t) + \psi_0(w_1(t)) + \int_0^t (-\phi_B(w_2(s))) \, ds \geq 0. \quad (3.6)$$

Then (3.5)-(3.6) has a unique solution $w \in D^2$ so that there is a well defined function $\xi : \mathbb{R}_+ \times \mathbb{R}^2 \times D^2 \to D^2$ mapping $(B, b, y)$ into $w = \xi(B, b, y)$. Furthermore, the function $\xi$ is continuous with respect to the product topology, with $\mathbb{R}_+$ equipped with the order topology and $D$ equipped with the topology of uniform convergence over bounded intervals. Finally, if $y$ is continuous, then so is $w$.

Before proceeding with the proof we introduce a version of Gronwall’s inequality first proved by Greene \[7\] and proved in the form we use by Das \[3\]:

**Lemma 3** (Gronwall’s inequality). Let $K_1$ and $K_2$ be nonnegative constants, let $h_i$ be real constants, and let $f, g$ be continuous nonnegative functions for all $t \geq 0$ such that

$$f(t) \leq K_1 + h_1 \int_0^t f(s) \, ds + h_2 \int_0^t g(s) \, ds,$$

$$g(t) \leq K_2 + h_3 \int_0^t f(s) \, ds + h_4 \int_0^t g(s) \, ds$$

for all $t \geq 0$. Then

$$f(t) \leq Me^{ht} \quad \text{and} \quad g(t) \leq Me^{ht}$$

for all $t \geq 0$ where $M = K_1 + K_2$ and $h = \max\{h_1 + h_3, h_2 + h_4\}$. In particular, if $K_1, K_2 = 0$, then $f(t), g(t) = 0$ for all $t$.

**Proof of Lemma 3**. We will show existence via a contraction mapping argument. First we will show that for $t \geq 0$ there exists a solution $\tilde{w} = (\tilde{w}_1, \tilde{w}_2)$ to the system of integral equations

$$\tilde{w}_1(t) = b_1 + y_1(t) + \int_0^t (-\phi_0(\tilde{w}_1(s)) + \phi_B(\tilde{w}_2(s) + \psi_0(\tilde{w}_1(s)))) \, ds, \quad (3.7)$$

$$\tilde{w}_2(t) = b_2 + y_2(t) + \int_0^t (-\phi_B(\tilde{w}_2(s) + \psi_0(\tilde{w}_1(s)))) \, ds \geq 0. \quad (3.8)$$
Once we have such a solution, it follows immediately that
\[ w = (w_1, w_2) = (\tilde{w}_1, \tilde{w}_2 + \psi_0(\tilde{w}_1)) \]
is a solution to (3.5)-(3.6).

We first show that the map defined by the right hand side of (3.7)-(3.8) is a contraction for small enough \( t \). We define \( T : D^2 \to D^2 \) by
\[
T(\tilde{w})_1(t) = b_1 + y_1(t) + \int_0^t (-\phi_0(\tilde{w}_1(s)) + \phi_B(\tilde{w}_2(s) + \psi_0(\tilde{w}_1(s)))) \, ds,
\]
\[
T(\tilde{w})_2(t) = b_2 + y_2(t) + \int_0^t (-\phi_B(\tilde{w}_2(s) + \psi_0(\tilde{w}_1(s)))) \, ds.
\]

For \( \tilde{w}, \tilde{v} \in D^2 \) we have
\[
||T(\tilde{w})_1 - T(\tilde{v})_1||_t \leq \int_0^t ||-\phi_0(\tilde{w}_1) + \phi_0(\tilde{v}_1)||_s \, ds
\]
\[
+ \int_0^t ||\phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1)) - \phi_B(\tilde{v}_2 + \psi_0(\tilde{v}_1))||_s \, ds
\]
\[ \leq 2 \int_0^t ||\tilde{w}_1 - \tilde{v}_1||_s \, ds
\]
\[ + 2 \int_0^t ||\tilde{w}_2 + \psi_0(\tilde{w}_1) - \tilde{v}_2 - \psi_0(\tilde{v}_1)||_s \, ds
\]
\[ \leq 2t ||\tilde{w}_1 - \tilde{v}_1||_t + 2t ||\tilde{w}_2 - \tilde{v}_2||_t + \int_0^t ||\tilde{w}_1 - \tilde{v}_1||_s \, ds
\]
\[ \leq 2t ||\tilde{w}_1 - \tilde{v}_1||_t + 2t ||\tilde{w}_2 - \tilde{v}_2||_t + t ||\tilde{w}_1 - \tilde{v}_1||_t
\]
\[ \leq 5t ||\tilde{w} - \tilde{v}||_t
\]
and
\[
||\tilde{w}_2 - \tilde{v}_2||_t \leq \int_0^t ||-\phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1)) + \phi_B(\tilde{v}_2 + \psi_0(\tilde{v}_1))||_s \, ds
\]
\[ \leq 2t ||\tilde{w}_2 + \psi_0(\tilde{w}_1) - \tilde{v}_2 - \psi_0(\tilde{v}_1)||_t
\]
\[ \leq 2t ||\tilde{w}_2 - \tilde{v}_2||_t + t ||\tilde{w}_1 - \tilde{v}_1||_t
\]
\[ \leq 3t ||\tilde{w} - \tilde{v}||_t.
\]

We therefore conclude that
\[ ||T(\tilde{w}) - T(\tilde{v})||_t \leq 5t ||\tilde{w} - \tilde{v}||_t, \]
so for \( t_0 < \frac{1}{5} \), \( T \) is a contraction on \( D([0, t_0], \mathbb{R}^2) \). Therefore by the contraction mapping principle (see, e.g., [15], p.220), \( T \) has a unique fixed point \( \tilde{w} \) on \( D([0, t_0], \mathbb{R}^2) \) such that \( T(\tilde{w}) = \tilde{w} \). This fixed point solves (3.7)-(3.8) for \( t \in [0, t_0] \). Now we extend the fixed point argument to \( t \in [t_0, 2t_0], [2t_0, 3t_0], \ldots \) and repeat to find a solution \( \tilde{w} \) to (3.7)-(3.8) for \( t \geq 0 \). As noted above, this provides a solution \( w \) to (3.5)-(3.6).

To prove uniqueness of this solution, suppose \( w \) and \( w' \) are two solutions to (3.5)-(3.6). We consider
\[
||w_1 - w'_1||_t \leq \int_0^t ||-\phi_0(w_1) + \phi_0(w'_1) + \phi_B(w_2) - \phi_B(w'_2)||_s \, ds
\]
\[ \leq 2 \int_0^t (||w_1 - w'_1||_s + ||w_2 - w'_2||_s) \, ds
\] (3.11)
\[\begin{align*}
|w_2 - w'_2|_t & \leq |\psi_0(w_1) - \psi_0(w'_1)|_t + \int_0^t |\phi_B(w_2) - \phi_B(w'_2)|_s \, ds \\
& \leq |w_1 - w'_1|_t + 2 \int_0^t |w_2 - w'_2|_s \, ds.
\end{align*}\] (3.12)

To match the form of Gronwall’s inequality (Lemma 3) we rewrite (3.12) as
\[\begin{align*}
|w_2 - w'_2|_t - |w_1 - w'_1|_t & \leq 2 \int_0^t |w_2 - w'_2|_s \, ds
\end{align*}\]
and note that the right hand side is nonnegative so the inequality remains true as
\[\begin{align*}
(|w_2 - w'_2|_t - |w_1 - w'_1|_t)^+ & \leq 2 \int_0^t |w_2 - w'_2|_s \, ds
\end{align*}\] (3.13)

We now define
\[u_1(t) = |w_1 - w'_1|_t,\]
\[u_2(t) = (|w_2 - w'_2|_t - |w_1 - w'_1|_t)^+\]
and note
\[|w_2 - w'_2|_s \leq u_2(s) + u_1(s) \quad s \geq 0.\] (3.14)

Then (3.11), (3.13), and (3.14) imply
\[\begin{align*}
u_1(t) & \leq 4 \int_0^t u_1(s) \, ds + 2 \int_0^t u_2(s) \, ds, \\
u_2(t) & \leq 2 \int_0^t u_1(s) \, ds + 2 \int_0^t u_2(s) \, ds.
\end{align*}\]

Now by Gronwall’s inequality we have
\[u_1(t) = 0 \quad \text{and} \quad u_2(t) = 0,\]
so by the definition of \(u_1\) and (3.14) we have
\[|w_1 - w'_1|_t = |w_2 - w'_2|_t = 0\]
for all \(t \geq 0\) and therefore the solution \(w\) is unique.

We now establish the continuity of \(\xi\). Suppose
\[(B^n, b^n, y^n) \to (B, b, y) \quad \text{as} \quad n \to \infty.\]
Fix \(\epsilon > 0\) and suppose \(w^n\) and \(w\) satisfy (3.5)–(3.6) for \((B^n, b^n, y^n)\) and \((B, b, y)\), respectively. Choose \(N\) such that for all \(n \geq N\)
\[|b^n - b| + ||y^n - y||_t + ||\phi_B^n(w_2) - \phi_B(w_2)||_t < \delta\]
for some \(\delta > 0\) which is yet to be determined. Note that such an \(N\) exists by Lemma 1 and the
assumption $B^n \to B$. We have
\[
||w^n_1 - w_1||_t \leq ||b^n - b|| + ||y^n - y||_t \\
+ \int_0^t ||-\phi_0(w^n_1) + \phi_0(w_1) + \phi_B(w^n_0 - \phi_B(w_2)||_s ds \\
\leq \delta + \int_0^t (2||w^n_1 - w_1||_s + ||\phi_B(w^n_0 - \phi_B(w_2)||_s \\
+ ||\phi_B(w_2) - \phi_B(w_2)||_s) ds \\
\leq \delta + \int_0^t (2||w^n_1 - w_1||_s + 2||w^n_2 - w_2||_s + \delta) ds \\
\leq \delta(1 + t) + 2\int_0^t (||w^n_1 - w_1||_s + ||w^n_2 - w_2||_s) ds \\
(3.15)
\]
and
\[
||w^n_2 - w_2||_t \leq \delta(1 + t) + ||w^n_1 - w_1||_t + 2\int_0^t ||w^n_2 - w_2||_s ds. \\
(3.16)
\]
As in the uniqueness argument above, we will apply Gronwall’s inequality, with functions
\[
u_1(t) = ||w^n_1 - w_1||_t \quad \text{and} \quad u_2(t) = (||w^n_2 - w_2||_t - ||w^n_1 - w_1||_t)^+.
\]
Then we have
\[
u_1(t) \leq \delta(1 + t) + 4\int_0^t u_1(s) ds + 2\int_0^t u_2(s) ds, \\
u_2(t) \leq \delta(1 + t) + 2\int_0^t u_1(s) ds + 2\int_0^t u_2(s) ds,
\]
so Gronwall’s inequality implies
\[
u_1(t) \leq 2\delta(1 + t)e^{6t} \quad \text{and} \quad u_2(t) \leq 2\delta(1 + t)e^{6t}
\]
and we have
\[
||w^n_1 - w_1||_t \leq 2\delta(1 + t)e^{6t} \quad \text{and} \quad ||w^n_2 - w_2||_t \leq 4\delta(1 + t)e^{6t}.
\]
We choose $\delta = \frac{1}{44}e^{-6t}$ to establish the desired continuity.

For the proof of continuity of $w$ we note
\[
|w_1(t + s) - w_1(t)| \leq |y_1(t + s) - y_1(t)| + \int_t^{t+s} |\phi_B(w_2(z)) - \phi_0(w_1(z))|dz
\]
and
\[
|w_2(t + s) - w_2(t)| \leq |y_2(t + s) - y_2(t)| + |w_1(t + s) - w_1(t)| + \int_t^{t+s} |\phi_B(w_2(z))|dz
\]
The boundedness of $w_1$ and $w_2$ proved in Lemma 7 imply that $w_1$ and $w_2$ are continuous if $y_1$ and $y_2$ are continuous.

We are now prepared to prove Theorem 4.
Proof of Theorem 1. Our key insight is to see that a solution is found by setting \( x_1 = \phi_0(w_1) \), \( u_1 = \psi_0(w_1) \), \( x_2 = \phi_B(w_2) \), and \( u_2 = \psi_B(w_2) \) where \((w_1, w_2)\) is the unique solution defined by Lemma 2.

To see that it is unique, note that the conditions on \( u_1 \) and \( u_2 \) imply that they can be written as \( \psi_0(z_1) \) and \( \psi_B(z_2) \) for some functions \( z_1, z_2 \in D \). Then \( x_1 \) and \( x_2 \) are \( \phi_0(z_1) \) and \( \phi_B(z_2) \) for the same \( z_1 \) and \( z_2 \). Then (2.4)-(2.6) imply that \( z = (z_1, z_2) \) must be a solution of (3.5)-(3.6). By Lemma 2 this solution is unique. In particular, this solution is as a Poisson process with rate equal to the number of customers in service. Since \( Q_1 \) then includes queues of length 1 and length 2, however, \( Q_1 \) will only decrease when a customer departs a queue and leaves the server empty. Therefore the instantaneous rate at time \( \lambda \) then arrivals are counted at their full rate.

4 Martingale representation.

We will now construct the process \( X^n(t) \) and show that it has the integral form in Theorem 1. Our representation will be similar to the first martingale representation of [14]; in particular it will rely upon random time changes of rate-1 Poisson processes.

4.1 Random time change.

We let \( A, D_1, \) and \( D_2 \) be rate-1 Poisson processes and write

\[
Q^n_1(t) = Q^n_1(0) + A(\lambda n t) - D_1 \left( \int_0^t (Q^n_1(s) - Q^n_2(s)) \, ds \right) - U^n_1(t),
\]

\[
Q^n_2(t) = Q^n_2(0) + U^n_1(t) - D_2 \left( \int_0^t Q^n_2(s) \, ds \right) - U^n_2(t),
\]

where \( U^n_1(t) \) is the number of arrivals in \([0, t]\) when every server has at least one customer, and \( U^n_2(t) \) is the number of arrivals in \([0, t]\) when every server has at least one customer and exactly \( B \sqrt{n} \) servers have two customers. Note that for this definition to make sense we must have \( B \leq \sqrt{n} \). Formally, we define

\[
U^n_1(t) = \int_0^t \mathbb{1} \{ Q^n_1(s) = n \} \, dA(\lambda n s),
\]

\[
U^n_2(t) = \int_0^t \mathbb{1} \{ Q^n_1(s) = n, Q^n_2(s) = B \sqrt{n} \} \, dA(\lambda n s).
\]

We can understand (4.1) term-by-term: first we record the initial state of the system with \( Q^n_1(0) \), then arrivals are counted at their full rate \( \lambda n \). The \( D_1 \) term represents departures, which occur as a Poisson process with rate equal to the number of customers in service. Since \( Q_1 \) includes queues of length 1 and length 2, however, \( Q_1 \) will only decrease when a customer departs a queue and leaves the server empty. Therefore the instantaneous rate at time \( s \) in the \( D_1 \) term is \( Q^n_1(s) - Q^n_2(s) \), the number of queues of length exactly 1 at time \( s \). Through the first three terms of (4.1) we have recorded what the value of \( Q_1 \) would be if it were not constrained to be at most \( n \), so the final term will represent this barrier. The process \( U^n_1 \) records any arrival which would increase \( Q_1 \) above \( n \), balancing the overcounting we get from \( A(\lambda n t) \).
We can understand (4.2) in much the same way, with the key difference being in the arrival process. Since arriving customers will always join the shortest available queue, the number of length 2 queues will increase only when all servers are busy. Such arrivals are exactly recorded by \( U^n_2 \), so this will be the process we use to record potential increases to \( Q^n_2 \). The process \( U^n_2 \) provides the upper barrier \( B \sqrt{n} \) on \( Q^n_2 \). We note that this regulating process must incorporate information about \( Q^n_1 \) as well as \( Q^n_2 \) because of the way increases to \( Q^n_2 \) rely on the value of \( Q^n_1 \). Therefore the arrivals that need to be “rejected” from the \( Q^n_2 \) process are precisely those that occur when \( Q^n_1(s) = n \) and \( Q^n_2(s) = B \sqrt{n} \), since the uncompensated result of such an arrival would be \( Q^n_2(s) > B \sqrt{n} \). We will see in Section 4.2, however, that this dependence on \( Q^n_1 \) can be hidden through properties of the one-dimensional reflection map.

As in [14, Lemma 2.1], we can verify that this construction is well defined and generates an element of \( D^2 \) by conditioning on the starting state \( Q^n(0) \) and processes \( A, D_1, D_2 \) then constructing recursively.

### 4.2 Martingales.

Because our approach to (4.1)-(4.2) will be to apply the functional central limit theorem (FCLT) for Poisson processes, we will now rewrite the time changes of Poisson processes as time changes of scaled Poisson processes. To that end, we define scaled martingales

\[
M^{n,1}_1(t) = \frac{1}{\sqrt{n}} A(\lambda_n nt) - \lambda_n \sqrt{n} t, \\
M^{n,2}_1(t) = \frac{1}{\sqrt{n}} D_1 \left( \int_0^t (Q^n_1(s) - Q^n_2(s)) ds \right) \\
- \frac{1}{\sqrt{n}} \int_0^t (Q^n_1(s) - Q^n_2(s)) ds, \\
M^{n,2}_2(t) = \frac{1}{\sqrt{n}} D_2 \left( \int_0^t Q^n_2(s) ds \right) - \frac{1}{\sqrt{n}} \int_0^t Q^n_2(s) ds,
\]

and also define

\[
V^n_1(t) = \frac{U^n_1(t)}{\sqrt{n}} \quad \text{and} \quad V^n_2(t) = \frac{U^n_2(t)}{\sqrt{n}}.
\]

Then we have

\[
X^n_2(t) = \frac{Q^n_1(t) - n}{\sqrt{n}} \\
= \frac{Q^n_1(0) - n}{\sqrt{n}} + \frac{1}{\sqrt{n}} A(\lambda_n nt) \\
- \frac{1}{\sqrt{n}} D_1 \left( \int_0^t (Q^n_1(s) - Q^n_2(s)) ds \right) - \frac{U^n_1(t)}{\sqrt{n}} \\
= X^n_2(0) + M^{n,1}_1(t) + \lambda_n \sqrt{n} t - V^n_1(t) \\
- M^{n,2}_1(t) - \frac{1}{\sqrt{n}} \int_0^t (Q^n_1(s) - Q^n_2(s)) ds \\
= X^n_2(0) + M^{n,1}_1(t) - M^{n,2}_1(t) + \lambda_n \sqrt{n} t - V^n_1(t) \\
- \sqrt{n} t - \int_0^t \left( \frac{Q^n_1(s) - n}{\sqrt{n}} - \frac{Q^n_2(s)}{\sqrt{n}} \right) ds \\
= X^n_2(0) + M^{n,1}_1(t) - M^{n,2}_1(t) - (1 - \lambda_n) \sqrt{n} t \\
- \int_0^t (X^n_1(s) - X^n_2(s)) ds - V^n_1(t)
\]
and

\[ X^n_2(t) = X^n_2(0) + V^n_1(t) - M^{n,2}_2(t) - \int_0^t X^n_2(s)ds - V^n_2(t). \tag{4.6} \]

Via an argument exactly analogous to that of in §7.1 of [14] leading to Theorem 7.2 we obtain that \( M^{n,1}_1, M^{n,2}_1, \) and \( M^{n,2}_2 \) are square-integrable martingales with respect to an appropriate filtration. We note for later use that this argument also supplies the predictable quadratic variations

\[ \langle M^{n,1}_1 \rangle(t) = \lambda_n t, \tag{4.7} \]
\[ \langle M^{n,2}_1 \rangle(t) = \frac{1}{n} \int_0^t (Q^n_1(s) - Q^n_2(s))ds, \tag{4.8} \]
\[ \langle M^{n,2}_2 \rangle(t) = \frac{1}{n} \int_0^t Q^n_2(s)ds. \tag{4.9} \]

At this point we can also note that (4.5)-(4.6) put \( X^n(t) \) in the integral form of Theorem 1. The only potential differences are the processes \( V^n \), which are not described in exactly the same way. We see, however, that by (4.3) we have

\[ 0 = \int_0^\infty \mathbb{1}\{Q^n_1(s) < n\}dU^n_1(s) \]
\[ = \int_0^\infty \mathbb{1}\{X^n_1(s) < 0\}dU^n_1(s) \]
\[ = \int_0^\infty \mathbb{1}\{X^n_1(s) < 0\}dV^n_1(s). \tag{4.10} \]

Similarly by (4.4) we have

\[ 0 = \int_0^\infty \mathbb{1}\{Q^n_2(s) < n \text{ or } Q^n_2(s) < B\sqrt{n}\}dU^n_2(t). \]

Because our definition of \( U^n_2 \) uses information about \( Q^n_1 \) this is the natural way to write a condition about the relationship between increases in \( U^n_2 \) and the values of \( Q^n \). We notice however that \( U^n_2 \) can increase only when we have both \( Q^n_1(s) = n \) and \( Q^n_2(s) = B\sqrt{n} \). In particular this means that it cannot increase when \( Q^n_2(s) < B\sqrt{n} \) regardless of the value of \( Q^n_1(s) \), so we also have the condition

\[ 0 = \int_0^\infty \mathbb{1}\{Q^n_2(s) < B\sqrt{n}\}dU^n_2(t) \]
\[ = \int_0^\infty \mathbb{1}\{X^n_2(s) < B\}dV^n_2(t). \tag{4.11} \]

Note that here we are “hiding” the dependence of \( U^n_2 \) on \( Q^n_1 \). Intuitively we are able to do this because the dependence on \( Q^n_1 \) comes entirely through the fact that increases to \( Q^n_2 \) are from \( U^n_1 \), which is captured in the integral representation itself. \( U^n_2 \) is acting only as a regulator for the reflecting upper barrier on \( Q^n_2 \), so it can be described as a simple reflecting barrier.

Based on (4.5)-(4.6) and (4.10)-(4.11) we have \( X^n \) and \( X \) represented according to Theorem 1 so to apply the CMT it remains to prove the convergence of the martingale pieces of (4.5)-(4.6).

5 Martingale convergence.

We will now prove the convergence of \( M^{n,i}_k \) to Brownian motions. In particular we prove
Lemma 4. For the sequences of scaled martingales $M_k^{n,i}$ defined in Section 4.2 we have the convergence

\[ (M_1^{n,1}, M_1^{n,2}, M_2^{n,2}) \Rightarrow (W_1, W_2, 0) \quad \text{in } D \quad \text{as } n \to \infty, \]

where $W_1$ and $W_2$ are independent standard Brownian motions.

To prove this lemma we will rely upon the CMT and the FCLT for Poisson processes ([14, Theorem 4.2]), which we state for our purposes as

Lemma 5. (FCLT for independent Poisson processes) If $A$, $D_1$, and $D_2$ are independent rate-1 Poisson processes and

\[ M_{C,n}(t) = \frac{C(nt) - nt}{\sqrt{n}} \]

for $C = A, D_1, D_2$, then

\[ (M_{A,n}, M_{D_1,n}, M_{D_2,n}) \Rightarrow (W_1, W_2, W_3) \quad \text{in } D^3 \quad \text{as } n \to \infty \]

where $W_1$, $W_2$, and $W_3$ are independent standard Brownian motions.

This result is a special case of the FCLT for a renewal process, which is discussed in [19]. More directly it can be derived via an application of the martingale FCLT, as discussed in §8 of [14].

To apply Lemma 5 we will define random and deterministic time changes such that the martingales $M_k^{n,i}$ can be written as a composition of a time change and the scaled Poisson processes $M_{C,n}$. Specifically, let

\begin{align*}
\Phi_{A,n}(t) &= \lambda_n t, \quad (5.2) \\
\Phi_{D_1,n}(t) &= \frac{1}{n} \int_0^t Q_{1}^n(s) ds - \frac{1}{n} \int_0^t Q_{2}^n(s) ds, \quad (5.3) \\
\Phi_{D_2,n}(t) &= \frac{1}{n} \int_0^t Q_{2}^n(s) ds, \quad (5.4)
\end{align*}

so that we have

\[ M_1^{n,1} = M_{A,n} \circ \Phi_{A,n}, \quad M_1^{n,2} = M_{D_1,n} \circ \Phi_{D_1,n}, \quad M_2^{n,2} = M_{D_2,n} \circ \Phi_{D_2,n}. \]

To apply the CMT with the composition map $\circ$, we need to determine the limits of the time changes (5.2)-(5.4).

First we note that (2.1) implies $\lambda_n \to 1$, which in turn implies

\[ \Phi_{A,n} \Rightarrow \epsilon \quad \text{as } n \to \infty, \]

where $\epsilon$ is the identity function in $D$.

Next we note that the second term of (5.3) is precisely $\Phi_{D_2,n}$, so

\[ \Phi_{D_1,n} \Rightarrow f - g \quad \text{as } n \to \infty, \]

where $g$ is the limit of $\Phi_{D_2,n}$ and $f$ is the limit of $\Phi_{D_1,n}$ with

\[ \Phi_{D_1,n}(t) = \frac{1}{n} \int_0^t Q_{1}^n(s) ds. \]

To find $f$ and $g$ we will first show fluid limits for $Q_{1}^n$ and $Q_{2}^n$. 
Lemma 6. Let $\Psi_i^n$ for $i = 1, 2$ be defined by

$$\Psi_i^n(t) = \frac{Q_i^n(t)}{n}, \quad t \geq 0.$$ 

Then

$$\Psi_1^n \Rightarrow \omega \quad \text{and} \quad \Psi_2^n \Rightarrow 0 \quad \text{as} \quad n \to \infty$$

(5.6)

where $\omega(t) = 1$ for $t \geq 0$.

The proof of this lemma is found in Section 5.1. To use Lemma 6 we define a continuous function $h : D \to D$ by

$$h(x)(t) = \int_0^t x(s)ds$$

for $t \geq 0$. The $\Phi_{D_1,n} = h \circ \Psi_n$ so by the CMT and Lemma 6 we know $f = h \circ \omega$. Namely,

$$f(t) = \int_0^t 1ds = t$$

so $f = e$ is the identity function in $D$. Therefore we have

$$\Phi_{D_1,n} \Rightarrow e \quad \text{as} \quad n \to \infty.$$ 

Similarly we have $g(t) = \int_0^t 0ds = 0$ so $g = 0$ on $D$. We conclude that

$$\Phi_{D_1,n} \Rightarrow e \quad \text{as} \quad n \to \infty,$$

(5.7)

and

$$\Phi_{D_2,n} \Rightarrow 0 \quad \text{as} \quad n \to \infty.$$ 

(5.8)

Therefore once we establish Lemma 6 we can prove Lemma 4.

Proof of Lemma 4. We apply the CMT with Lemma 5 and the limits (5.5), (5.7), and (5.8) to obtain

$$M_1^{n,1}, M_1^{n,2}, M_2^{n,2} = (M_{A,n}, M_{A,n} \circ \Phi_{D_1,n}, M_{D_2,n} \circ \Phi_{D_2,n})$$

$$\Rightarrow (W_1 \circ e, W_2 \circ e, W_3 \circ 0)$$

$$= (W_1, W_2, 0).$$

5.1 Fluid limit.

We will prove Lemma 6 by showing that $X^n_1$ is stochastically bounded in $D$. Namely, we will prove that the sequence of real-valued random variables $\|X^n_t\|_1$ is tight for every $t > 0$. For a more complete discussion of stochastic boundedness as we use it here see §5 of [14].

The stochastic boundedness of $X^n_1$ will follow from the stochastic boundedness of $M_1^{n,1}$, $M_1^{n,2}$, and $M_2^{n,2}$. To see this, we prove
Lemma 7. Given \((B_n, X^n_i(0), Y^n_i)\) a random element of \(\mathbb{R}_+ \times \mathbb{R} \times D\) for each \(n \geq 1\) and \(i = 1, 2\), recall that Theorem 1 implies that the system

\[
X^n_1(t) = X^n_1(0) + Y^n_1(t) + \int_0^t (-X^n_1(s) + X^n_2(s))ds - V^n_1(t),
\]

\[
X^n_2(t) = X^n_2(0) + Y^n_2(t) + \int_0^t (-X^n_2(s))ds + V^n_1(t) - V^n_2(t) \geq 0,
\]

\[
0 = \int_0^\infty 1\{X^n_i(t) < 0\}dV^n_i(t),
\]

\[
0 = \int_0^\infty 1\{X^n_i(t) < B_n\}dV^n_i(t).
\]

has a unique solution \((X^n, V^n)\). If the sequences \((X^n(0), n \geq 1)\) and \((Y^n, n \geq 1)\) are stochastically bounded for \(i = 1, 2\), then the sequence \((X^n, n \geq 1)\) is stochastically bounded in \(D\).

Note that we do not require boundedness for \(B_n\).

Proof. We fix \(t > 0\). We will establish the bound

\[
||X^n||_t \leq 8e^{6t} (||X^n(0)|| + ||Y^n||_t),
\]

from which the result follows.

To show \(5.9\), we will prove a similar bound for the unreflected process \(W^n\) defined by Lemma 2. Then \(5.9\) will follow from the Lipschitz continuity of the reflection maps \(\phi_0\) and \(\phi_{B_n}\).

Just as in Theorem 1 and Lemma 2, we write \(X^n_1(t) = \phi_0(W^n_1(t))\) and \(X^n_2(t) = \phi_{B^n}(W^n_2(t))\) where \(W^n_1(t)\) and \(W^n_2(t)\) satisfy

\[
W^n_1(t) = X^n_1(0) + Y^n_1(t) + \int_0^t (-\phi_0(W^n_1(s)) + \phi_{B^n}(W^n_2(s)))ds,
\]

\[
W^n_2(t) = X^n_2(0) + Y^n_2(t) + \int_0^t (-\phi_{B^n}(W^n_2(s)))ds + \psi_0(W^n_1(t)).
\]

We now use Gronwall’s inequality as stated in Lemma 3. Using the Lipschitz property for \(\phi_0, \phi_{B_n}\), and \(\psi_0\) we have for \(t \geq 0\)

\[
||W^n_1||_t \leq ||X^n_1(0)|| + ||Y^n||_t + 2\int_0^t (||W^n_2||_s + ||W^n_1||_s)ds,
\]

\[
||W^n_2||_t \leq ||X^n_2(0)|| + ||Y^n||_t + ||\psi_0(W^n_1)||_t + \int_0^t ||W^n_2||_s ds.
\]

Now we note that we have

\[
||\psi_0(W^n_1)||_t \leq ||W^n_1||_t.
\]

We define

\[
u_1(t) = ||W^n_1||_t \quad \text{and} \quad u_2(t) = (||W^n_2||_t - ||W^n_1||_t)^+.
\]

Finally we note

\[
||W^n_2||_t \leq u_2(t) + u_1(t),
\]

so we can write the inequalities

\[
u_1(t) \leq |X^n_1(0)| + ||Y^n||_t + 4\int_0^t u_1(s)ds + 2\int_0^t u_2(s)ds,
\]

\[
u_2(t) \leq |X^n_2(0)| + ||Y^n||_t + \int_0^t u_1(s)ds + \int_0^t u_2(s)ds.
\]
Let $|X^n(0)| + ||Y^n||_t = K$. Then Lemma 3 implies

$$u_1(t) \leq 2Ke^{6t} \text{ and } u_2(t) \leq 2Ke^{6t}.$$ 

From (5.12) and the definitions of $u_1$ and $u_2$ we obtain

$$||W^n_1||_t \leq 2Ke^{6t} \text{ and } ||W^n_2||_t \leq 4Ke^{6t}.$$ 

Since $\phi_0$ and $\phi_B$ are Lipschitz continuous with constant 2 this implies

$$||X^n_1||_t \leq 4Ke^{6t} \text{ and } ||X^n_2||_t \leq 8Ke^{6t},$$ 

which proves (5.9).

Note that this proof also provides the boundedness of $w$ that we use in the proof of continuity of $w$ in Lemma 2 and that it does not use any of the continuity properties proved using that boundedness.

In our application of Lemma 7 we will have $Y^n_1 = M^n_1 - M^n_2 - (1-\lambda^n)\sqrt{nt}$ and $Y^n_2 = M^n_2$, so it remains to prove that each martingale $M^n_{k,i}$ is stochastically bounded. To prove the stochastic boundedness of these martingales we will use the following lemma from [14]:

**Lemma 8** ([14] Lemma 5.8). Suppose that, for each $n \geq 1$, $M_n$ is a square integrable martingale with predictable quadratic variation $\langle M_n \rangle$. If the sequence of random variables $\langle M_n \rangle(T)$ is stochastically bounded in $\mathbb{R}$ for each $T > 0$, then the sequence of stochastic processes $M_n$ is stochastically bounded in $D$.

We now prove that the predictable quadratic variations of $M^n_{k,i}$ are stochastically bounded. In the case of $M^n_{1,1}$ this is immediate since by (4.7) the quadratic variation is deterministic.

For $M^n_{1,2}$ we refer to (4.8) and apply crude bounds to see

$$\langle M^n_{1,2} \rangle(t) = \frac{1}{n} \int_0^t (Q^n_1(s) - Q^n_2(s))ds$$

$$\leq \frac{1}{n} \int_0^t Q^n_1(s)ds$$

$$\leq \frac{t}{n} (Q^n_1(0) + A(\lambda, nt)).$$

It suffices to show stochastic boundedness of each term in the sum. For $Q^n_1(0)$ this follows from assumption (2.7).

For $A(\lambda, nt)$ we note $\lambda_n \to 1$ so by the strong law of large numbers (SLLN) for Poisson processes we have

$$\frac{A(\lambda, nt)}{n} \to e(t)$$

with probability 1, which implies stochastic boundedness, so we conclude that $M^n_{1,2}$ is stochastically bounded.

For $M^n_{2,2}$ we have

$$\langle M^n_{2,2} \rangle(t) \leq \frac{t}{n} (Q^n_2(0) + U^n_1(t))$$

$$\leq \frac{t}{n} (Q^n_2(0) + A(\lambda, nt)),$$

and stochastic boundedness follows.

We now return to the proof of Lemma 0.
Proof of Lemma 6. We have that

\[ X^n_1 = Q^n_1 - \frac{n}{\sqrt{n}} \quad \text{and} \quad X^n_2 = \frac{Q^n_2}{\sqrt{n}} \]

are stochastically bounded. Therefore

\[ \frac{X^n_i}{\sqrt{n}} \Rightarrow 0 \quad \text{in} \quad D \quad \text{as} \quad n \to \infty. \]

From the definition of \( X^n \) this is equivalent to

\[ \Psi^n_1 = \frac{Q^n_1}{n} \Rightarrow \omega \quad \text{and} \quad \Psi^n_2 = \frac{Q^n_2}{n} \Rightarrow 0 \quad \text{in} \quad D \quad \text{as} \quad n \to \infty. \]

\[ \blacksquare \]

5.2 Proof of Theorem 2.

Now that we have the convergence of the martingale processes \( M^{n,i}_k \), we can apply the CMT to prove Theorem 2.

Proof of Theorem 2. In Theorem 1, in the pre-limit regime we set \( B_n = \sqrt{n} \), \( b_1 = X^n_1(0) \), \( b_2 = X^n_2(0) \),

\[ y_1(t) = M^{n,1}_1(t) - M^{n,2}_2(t) - (1 - \lambda_n)\sqrt{nt}, \]

and \( y_2(t) = -M^{n,2}_2(t) \). Equations (4.10) and (4.11) show that \( V^n_1 \) and \( V^n_2 \) are appropriately acting as \( u_1 \) and \( u_2 \) in the integral representation, so \( x_1(t) = X^n_1(t) \) and \( x_2(t) = X^n_2(t) \).

For application of the CMT we need only determine the limits of \( B \), \( b \) and \( y \).

We have \( B_n \to \infty \). Note that this, along with the proof of Theorem 1 from Lemma 2 and the definition of \( \psi_\infty \), implies that the process \( u_2 \) in the limiting system will be identically zero, as it is in the statement of Theorem 2.

By assumption we have

\[ X^n_k(0) \Rightarrow X_k(0), \]

so in the limiting system we have \( b_1 = X_1(0) \) and \( b_2 = X_2(0) \). Next we have by (2.1) and (5.1)

\[ M^{n,1}_1(t) - M^{n,2}_2(t) - (1 - \lambda_n)\sqrt{nt} \Rightarrow W_1(t) - W_2(t) - \beta t \]

\[ \overset{d}{=} \sqrt{2}W(t) - \beta t, \]

where \( W \) is a standard Brownian motion and \( \overset{d}{=} \) indicates equivalence in distribution. Another application of (5.1) implies

\[ -M^{n,2}_2 \Rightarrow 0 \]

so in the limiting system we have \( y_1(t) = \sqrt{2}W(t) - \beta t \) and \( y_2(t) = 0 \).

The CMT then implies \( X^n_k \Rightarrow X_k \) in \( D \) as \( n \to \infty \) where \( X_k \) is described by (2.8)-(2.10). \[ \blacksquare \]

6 Open questions.

Theorem 2 proves that the behavior of the truncated \( M/M/n \)-JSQ system in the Halfin-Whitt regime is best understood on the order of \( O(\sqrt{n}) \). In particular, the numbers of idle servers and waiting customers will both be \( O(\sqrt{n}) \).

Several open questions remain. The conjecture that this truncated system is a good model for the original untruncated system remains to be proven. Thus an extension of the result for the original model may be of interest.
Significant questions also remain about the steady state behavior of our system. In particular, we do not characterize the distribution of the steady state of the limiting system or show that the steady state of the $n$-th system converges to the steady state of the limiting diffusion process (interchange of limits).

Questions of the steady state are closely related to the waiting time for the system. Because customers immediately enter service if there are any idle servers and otherwise wait a constant order amount of time for the previous customer in their queue to finish service, the waiting time can be characterized by the amount of time when the system has no idle servers. We conjecture that this time is on the order $O(1/\sqrt{n})$, which would lead to an expected waiting time of the same order. This is also the order of the expected waiting time in the $M/M/n$ system, so a proof of this conjecture would demonstrate that the JSQ system has a minimal loss of efficiency as measured by expected waiting time.

Finally it is always of interest to analyze our system for general interarrival and, especially, general service times distribution. We conjecture that the qualitative behaviour established in this paper in the transient domain and the conjectures above regarding the steady-state behavior and the interchange of steady-state limits remain true in this case as well.

### Acknowledgments.

This work was supported by NSF grant CMMI-1335155.

### References

[1] G. Brightwell and M. Luczak. The supermarket model with arrival rate tending to one. *ArXiv e-prints*, January 2012.

[2] J. G. Dai and Tolga Tezcan. State space collapse in many-server diffusion limits of parallel server systems. *Mathematics of Operations Research*, 36(2):pp. 271–320, 2011.

[3] K. M. Das. A note on an inequality due to Greene. *Proc. Amer. Math. Soc.*, 77(3):424–425, 1979.

[4] A. B. Dieker and T. Suk. Randomized longest-queue-first scheduling for large-scale buffered systems. *ArXiv e-prints*, June 2013.

[5] L. Flatto and H. P. McKean. Two queues in parallel. *Communications on Pure and Applied Mathematics*, 30(2):255–263, 1977.

[6] G.J. Foschini and J. Salz. A basic dynamic routing problem and diffusion. *Communications, IEEE Transactions on*, 26(3):320–327, Mar 1978.

[7] David E. Greene. An inequality for a class of integral systems. *Proceedings of the American Mathematical Society*, 62(1):pp. 101–104, 1977.

[8] Frank A. Haight. Two queues in parallel. *Biometrika*, 45(3-4):401–410, 1958.

[9] Shlomo Halfin. The shortest queue problem. *Journal of Applied Probability*, 22(4):pp. 865–878, 1985.

[10] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[11] Zhang Hanqin and Wang Rongxin. Heavy traffic limit theorems for a queueing system in which customers join the shortest line. *Advances in Applied Probability*, 21(2):pp. 451–469, 1989.

[12] J. F. C. Kingman. Two similar queues in parallel. *The Annals of Mathematical Statistics*, 32(4):1314–1323, 12 1961.
[13] M. Mitzenmacher. The power of two choices in randomized load balancing. *Parallel and Distributed Systems, IEEE Transactions on*, 12(10):1094–1104, 2001.

[14] Guodong Pang, Rishi Talreja, and Ward Whitt. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*, 4:193–267, 2007.

[15] Walter Rudin. *The Principles of Mathematical Analysis*. International Series in Pure & Applied Mathematics. McGraw-Hill Publishing Company, 3rd edition, 2006.

[16] Tolga Tezcan. Optimal control of distributed parallel server systems under the halfin and whitt regime. *Mathematics of Operations Research*, 33(1):51–90, 2008.

[17] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Probl. Peredachi Inf.*, 32(1):20–34, 1996.

[18] Richard R. Weber. On the optimal assignment of customers to parallel servers. *Journal of Applied Probability*, 15(2):pp. 406–413, 1978.

[19] Ward Whitt. *Stochastic-process Limits: An Introduction To Stochastic-process Limits And Their Application To Queues*. Springer Series In Operations Research. Springer, New York, 2002.

[20] Wayne Winston. Optimality of the Shortest Line Discipline. *Journal of Applied Probability*, 14(1):181–189, 1977.

[21] Hanquin Zhang, Guang-Hui Hsu, and Rongxin Wang. Heavy traffic limit theorems for a sequence of shortest queueing systems. *Queueing Systems*, 21(1-2):217–238, 1995.