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

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. We use the martingale method to prove that a scaled process counting the number of idle servers and queues of length exactly 2 weakly converges to a two-dimensional reflected Ornstein-Uhlenbeck process, while processes counting longer queues converge to a deterministic system decaying to zero in constant time. 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, only a vanishing fraction of customers will have to wait, but those who do will incur a constant order waiting 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 (i.e. $M/M/n$ and similar queues [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. To analyze this tradeoff, we will study 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 (Halfin-Whitt) heavy traffic by allowing the arrival rate λ_n to depend on n, letting the quantity $(1-\lambda_n)\sqrt{n}$ have a non-degenerate limit, which we 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 of queues with one customer waiting to enter service, along with auxiliary processes to count the number of longer queues. We prove that a system that initially has a fixed maximum queue length, appropriately scaled, converges weakly to a diffusion process as n approaches infinity. The coordinates of this diffusion process representing queues with more than 1 customer in service are deterministic and show that any longer queues present in the initial condition disappear in fixed time and do not form again. The coordinates corresponding to the number of idle servers and number of queues with exactly one customer waiting correspond to a two-dimensional reflected Ornstein-Uhlenbeck process. The entire limiting system 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.

As a proof technique, we will introduce a truncated variant of the $M/M/n$-JSQ system in which no queues of length longer than 2 are created, though such long queues are allowed in the initial condition. This system is more easily analyzed because it is finite dimensional and has limited interaction between many of the dimensions. In this truncated system, we show that the probability the system hits the truncation barrier decreases to zero as n approaches infinity, and thus in the limit the behavior of the truncated and untruncated systems are the same.

One consequence of our result is that both the number of idle servers and the number of queues with exactly one customer waiting in the $M/M/n$-JSQ system are of the order $O(\sqrt{n})$.

Another feature of interest in this queueing system is the waiting time experienced by arriving customers. We prove that in the transient system the aggregate waiting time experienced by all customers is of the order $O(\sqrt{n})$, and since the number of customers arriving in that time is order n, the waiting time per customer is $O(1/\sqrt{n})$. We also observe that any arriving customer who has to wait will incur a waiting time which is exponentially distributed with parameter 1, which is the service time of the customer in service when they enter a queue. Since any waiting customers must incur a constant order waiting time and the aggregate waiting time is $O(\sqrt{n})$, the fraction of customers who end up waiting is of the order $O(1/\sqrt{n})$.

Next we review prior 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.

Thus our paper is the first study of JSQ systems in the asymptotic regime as $n \to \infty$. Observe 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 $M/M/\infty$ system, and thus it is of interest to consider this model in heavy traffic with λ 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 λ 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 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 defines the model and states our main result. In Section 3 we 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 the convergence of the scaled queue length processes. This section will conclude our proof of Theorem 2. In Section 6 we discuss the waiting time in the $M/M/n$-JSQ system. We will conclude in Section 7 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 $\bar{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$ represent the extended positive real line. We will equip $\bar{\mathbb{R}}_+$ with the order topology, in which neighborhoods of ∞ are those sets which contain a subset of the form $\{x > a\}$ for some $a \in \mathbb{R}$. 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 and the main result.

We consider a $M/M/n$-JSQ queueing system with n servers where each server maintains a unique queue, with service proceeding according to the 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

$$\lim_{n\to\infty} \sqrt{n}(1-\lambda_n) = \beta \tag{2.1}$$

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 via the process $Q^n(t) = (Q_1^n(t), Q_2^n(t), \ldots)$, with $Q_i^n(t)$ representing the number of queues with *at least* i customers (including any customer in service) at time $t \geq 0$. We note that for the system as described we have

$$n \geq Q_1^n(t) \geq Q_2^n(t) \geq \cdots \geq 0, \qquad \forall t \geq 0, \tag{2.2}$$

and that we can recover the number of queues with exactly i customers in service via the quantity $Q_i^n(t) - Q_{i+1}^n(t)$, including the number of idle servers $n - Q_1^n(t)$.

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

$$X_1^n(t) = \frac{Q_1^n(t) - n}{\sqrt{n}} \quad \text{and} \quad X_i^n(t) = \frac{Q_i^n(t)}{\sqrt{n}}, \quad i \geq 2. \tag{2.3}$$

The $i = 1$ case is treated differently because the number of queues with length 1 behaves differently than the number of queues of all larger lengths. In particular, there will be $O(\sqrt{n})$ idle servers, and thus the number of servers with at least one customer in service will be order n.

Our diffusion limit will be the solution to a system of k integral equations for some $k \geq 3$, so we first introduce this system and prove that it has a unique solution. Furthermore, we prove that the system defines a continuous map from $\bar{\mathbb{R}}_+ \times \mathbb{R}^k \times D^k$ to $D^k \times D^2$ 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.

Theorem 1. *Given integer $k \geq 3$, $B \in \bar{\mathbb{R}}_+$, $b \in \mathbb{R}^k$, and $y \in D^k$, consider the following system:*

$$x_1(t) = b_1 + y_1(t) + \int_0^t (-x_1(s) + x_2(s))ds - u_1(t), \tag{2.4}$$

$$x_2(t) = b_2 + y_2(t) + \int_0^t (-x_2(s) + x_3(s))ds + u_1(t) - u_2(t), \tag{2.5}$$

$$x_i(t) = b_i + y_i(t) + \int_0^t (-x_i(s) + x_{i+1}(s))ds, \quad 3 \leq i \leq k-1, \tag{2.6}$$

$$x_k(t) = b_k + y_k(t) + \int_0^t -x_k(s)ds, \tag{2.7}$$

$$x_1(t) \leq 0, \quad 0 \leq x_2(t) \leq B, \quad x_i(t) \geq 0, \quad t \geq 0, \tag{2.8}$$

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)-(2.8) *has a unique solution* $(x,u) \in D^k \times D^2$ *so that there is a well defined function* $(f,g) : \bar{\mathbb{R}}_+ \times \mathbb{R}^k \times D^k \to D^k \times D^2$ *mapping* (B,b,y) *into* $x = f(B,b,y)$ *and* $u = g(B,b,y)$. *Furthermore, the function* (f,g) *is continuous on* $\bar{\mathbb{R}}_+ \times \mathbb{R}^k \times D^k$ *with respect to the product topology when* $\bar{\mathbb{R}}_+$ *is equipped with the order topology and* D *is equipped with the topology of uniform convergence over bounded intervals. 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 below is well defined because, as we will see, it is an application of the function (f,g) with specific arguments b, y, and $B = \infty$ augmented with $X_i(t) = 0$ for $i > k$. Note that $B = \infty$ implies $u_2 = 0$. Our main result is the following:

Theorem 2. *In the sequence of* $M/M/n$*-JSQ models described above, suppose there exists* k *and a random vector* $X(0) = (X_1(0), \ldots, X_k(0)) \in \mathbb{R}^k$ *such that*

$$X_i^n(0) \Rightarrow X_i(0) \quad \text{in } \mathbb{R} \text{ as } n \to \infty, \quad 1 \le i \le k, \tag{2.9}$$

and $X_i^n(0) = 0$ *for* $i > k$. *Then for any* $t \ge 0$,

$$X_i^n \Rightarrow X_i \quad \text{in } D \text{ as } n \to \infty, \quad i \ge 1,$$

where $X_1 \le 0$ *and* $X_i \ge 0$ *for* $i \ge 2$ *are unique solutions in* D *of the stochastic integral equations*

$$X_1(t) = X_1(0) + \sqrt{2}W(t) - \beta t + \int_0^t (-X_1(s) + X_2(s))\, ds - U_1(t), \tag{2.10}$$

$$X_2(t) = X_2(0) + U_1(t) + \int_0^t (-X_2(s) + X_3(s))ds, \tag{2.11}$$

$$X_i(t) = X_i(0) + \int_0^t (-X_i(s) + X_{i+1}(s))ds, \quad 3 \le i \le k-1, \tag{2.12}$$

$$X_k(t) = X_k(0) + \int_0^t -X_k(s)ds, \tag{2.13}$$

$$X_i(t) = 0, \quad i \ge k+1, \tag{2.14}$$

where W *is a standard Brownian motion and* U_1 *is the unique nondecreasing nonnegative process in* D *satisfying*

$$\int_0^\infty \mathbb{1}\{X_1(t) < 0\} dU_1(t) = 0. \tag{2.15}$$

Remark 1. The integral equations (2.12)-(2.13) are deterministic and have an explicit solution:

$$X_i(t) = e^{-t}\left(X_i(0) + \sum_{j=1}^{k-i} \frac{1}{j!} t^j X_{i+j}(0)\right), \quad 3 \le i \le k-1,$$
$$X_k(t) = X_k(0)e^{-t}.$$

Thus the number of queues of length at least i for $i \ge 3$ decays exponentially in time.

We note that condition (2.9) does place significant but not unreasonable restrictions on the starting state of the finite systems Q^n. In particular, $Q_1^n(0) - n = O(\sqrt{n})$ so the number of customers initially in service must be sufficiently near n. Similarly, (2.9) requires $Q_2^n(0) = O(\sqrt{n})$ and therefore $Q_i^n(0) = O(\sqrt{n})$ for $3 \le i \le k$. Also note that the longest queue allowed in the initial condition has length k.

Our result shows that the $M/M/n$-JSQ system in the heavy traffic limit becomes essentially a two-dimensional system. If queues with more than one customer waiting are present initially, they disappear and do not form again. There are $O(\sqrt{n})$ idle servers and $O(\sqrt{n})$ queues with exactly one customer, and the behavior of processes counting these correspond to a two-dimensional Ornstein-Uhlenbeck process.

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 $\bar{\mathbb{R}}_+ \times \mathbb{R}^k \times D^k$ to $D^k \times D^2$. 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 $X_i^n \Rightarrow X_i$ once we write X_i^n in the appropriate integral form.

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

Our approach to proving Theorem 1 will involve two main lemmas, one dealing with the first two dimensions, where reflection plays an important role, and a one dimensional lemma that we will apply to the higher dimensions.

3.1 Lower dimensions.

To deal with the reflection terms in the first two dimensions of Theorem 1 it is convenient to consider those dimensions completely decoupled from the rest of the system. To that end we will prove the following:

Lemma 1. *Given $B \in \bar{\mathbb{R}}_+$, $b \in \mathbb{R}^2$, and $y \in D^2$, consider*

$$x_1(t) = b_1 + y_1(t) + \int_0^t (-x_1(s) + x_2(s))ds - u_1(t), \tag{3.1}$$

$$x_2(t) = b_2 + y_2(t) + \int_0^t (-x_2(s))ds + u_1(t) - u_2(t), \tag{3.2}$$

$$x_1(t) \le 0, \quad 0 \le x_2(t) \le B, \quad t \ge 0, \tag{3.3}$$

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 (3.1)-(3.3) has a unique solution $(x, u) \in D^2 \times D^2$ so that there is a well defined function $(f, g) : \bar{\mathbb{R}}_+ \times \mathbb{R}^2 \times D^2 \to D^2 \times D^2$ mapping (B, b, y) into $x = f(B, b, y)$ and $u = g(B, b, y)$. Furthermore, the function (f, g) is continuous. Finally, if y is continuous, then so are x and u.

3.1.1 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 $(\phi_\kappa, \psi_\kappa) : D \to D^2$ be the one-sided reflection map with upper barrier at κ (see, e.g., [19] §5.2 and §13.5). In particular for $x \in D$ with $x(0) \le \kappa$ we have $z = \psi_\kappa(y) \ge 0$, z nondecreasing,

$$x = \phi_\kappa(y) = y - z \le \kappa,$$

and

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

Recall that these functions can be defined explicitly by

$$\psi_\kappa(x)(t) = \sup_{0 \le s \le t} (x(s) - \kappa)^+ \tag{3.4}$$

and

$$\phi_\kappa(x)(t) = x(t) - \psi_\kappa(x)(t). \tag{3.5}$$

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

$$||\psi_\kappa(x) - \psi_{\kappa'}(x')||_t \leq ||x - x'||_t + |\kappa - \kappa'|, \tag{3.6}$$

$$||\phi_\kappa(x) - \phi_{\kappa'}(x')||_t \leq 2\,||x - x'||_t + |\kappa - \kappa'|. \tag{3.7}$$

These follow straightforwardly from (3.4) and (3.5). Note that for $\kappa = \kappa'$ we recover the usual Lipschitz constants of 1 for ψ_κ and 2 for ϕ_κ.

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 2. *The function $(\phi, \psi) : \bar{\mathbb{R}}_+ \times D \to D^2$ defined by (3.4)-(3.5) for finite κ and by $(\phi_\infty, \psi_\infty) = (e, 0)$ for $\kappa = \infty$ is continuous with respect to the product topology when $\bar{\mathbb{R}}_+$ is equipped with the order topology and D is equipped with the topology of uniform convergence over bounded intervals.*

Proof. By (3.6)-(3.7) 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 $\kappa \to \infty$,

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

where we have made use of the fact that $||x||_t < \infty$. Therefore $\psi_\kappa(x^\kappa) \to \psi_\infty(x)$ and by (3.5) we conclude $\phi_\kappa(x^\kappa) \to \phi_\infty(x)$. Thus the function is continuous at $\kappa = \infty$, completing the proof. □

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

Lemma 3. *Given $B \in \bar{\mathbb{R}}_+$, $b \in \mathbb{R}^2$ and $y \in D^2$, consider*

$$w_1(t) = b_1 + y_1(t) + \int_0^t (-\phi_0(w_1(s)) + \phi_B(w_2(s)))\, ds, \tag{3.8}$$

$$w_2(t) = b_2 + y_2(t) + \psi_0(w_1(t)) + \int_0^t (-\phi_B(w_2(s)))\, ds \geq 0. \tag{3.9}$$

Then (3.8)-(3.9) has a unique solution $w \in D^2$ so that there is a well defined function $\xi : \bar{\mathbb{R}}_+ \times \mathbb{R}^2 \times D^2 \to D^2$ mapping (B, b, y) into $w = \xi(B, b, y)$. Furthermore, the function ξ is continuous. 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 4 (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

$$\tilde{w}_1(t) = b_1 + y_1(t) + \int_0^t \left(-\phi_0(\tilde{w}_1(s)) + \phi_B\big(\tilde{w}_2(s) + \psi_0(\tilde{w}_1(s))\big)\right) ds, \tag{3.10}$$

$$\tilde{w}_2(t) = b_2 + y_2(t) + \int_0^t \left(-\phi_B\big(\tilde{w}_2(s) + \psi_0(\tilde{w}_1(s))\big)\right) ds \geq 0. \tag{3.11}$$

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 (3.8)-(3.9).

We first show that the map defined by the right hand side of (3.10)-(3.11) is a contraction for small enough t. We define $T : D^2 \to D^2$ by

$$T(\tilde{w})_1(t) = b_1 + y_1(t) + \int_0^t \left(-\phi_0(\tilde{w}_1(s)) + \phi_B\big(\tilde{w}_2(s) + \psi_0(\tilde{w}_1(s))\big)\right) ds, \tag{3.12}$$

$$T(\tilde{w})_2(t) = b_2 + y_2(t) + \int_0^t \left(-\phi_B\big(\tilde{w}_2(s) + \psi_0(\tilde{w}_1(s))\big)\right) ds. \tag{3.13}$$

For $\tilde{w}, \tilde{v} \in D^2$ we have

$$\begin{aligned}
||T(\tilde{w})_1 - T(\tilde{v})_1||_t &\leq \int_0^t ||-\phi_0(\tilde{w}_1) + \phi_0(\tilde{v}_1)||_s \, ds \\
&\quad + \int_0^t ||\phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1)) - \phi_B(\tilde{v}_2 + \psi_0(\tilde{v}_1))||_s \, ds \\
&\leq 2 \int_0^t ||\tilde{w}_1 - \tilde{v}_1||_s \, ds \\
&\quad + 2 \int_0^t ||\tilde{w}_2 + \psi_0(\tilde{w}_1) - \tilde{v}_2 - \psi_0(\tilde{v}_1)||_s \, ds \\
&\leq 2t \, ||\tilde{w}_1 - \tilde{v}_1||_t + 2t \, ||\tilde{w}_2 - \tilde{v}_2||_t + \int_0^t ||\tilde{w}_1 - \tilde{v}_1||_s \, ds \\
&\leq 2t \, ||\tilde{w}_1 - \tilde{v}_1||_t + 2t \, ||\tilde{w}_2 - \tilde{v}_2||_t + t \, ||\tilde{w}_1 - \tilde{v}_1||_t \\
&\leq 5t \, ||\tilde{w} - \tilde{v}||_t
\end{aligned}$$

and

$$\begin{aligned}
||\tilde{w}_2 - \tilde{v}_2||_t &\leq \int_0^t ||-\phi_B(\tilde{w}_2 + \psi_0(\tilde{w}_1)) + \phi_B(\tilde{v}_2 + \psi_0(\tilde{v}_1))||_s \, ds \\
&\leq 2t \, ||\tilde{w}_2 + \psi_0(\tilde{w}_1) - \tilde{v}_2 - \psi_0(\tilde{v}_1)||_t \\
&\leq 2t \, ||\tilde{w}_2 - \tilde{v}_2||_t + t \, ||\tilde{w}_1 - \tilde{v}_1||_t \\
&\leq 3t \, ||\tilde{w} - \tilde{v}||_t .
\end{aligned}$$

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 (3.10)-(3.11) 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 (3.10)-(3.11) for $t \geq 0$. As noted above, this provides a solution w to (3.8)-(3.9).

To prove uniqueness of this solution, suppose w and w' are two solutions to (3.8)-(3.9). We consider

$$\begin{aligned} ||w_1 - w_1'||_t &\leq \int_0^t ||-\phi_0(w_1) + \phi_0(w_1') + \phi_B(w_2) - \phi_B(w_2')||_s \, ds \\ &\leq 2\int_0^t (||w_1 - w_1'||_s + ||w_2 - w_2'||_s) \, ds \end{aligned} \tag{3.14}$$

and

$$\begin{aligned} ||w_2 - w_2'||_t &\leq ||\psi_0(w_1) - \psi_0(w_1')||_t + \int_0^t ||\phi_B(w_2) - \phi_B(w_2')||_s \, ds \\ &\leq ||w_1 - w_1'||_t + 2\int_0^t ||w_2 - w_2'||_s \, ds. \end{aligned} \tag{3.15}$$

To match the form of Gronwall's inequality (Lemma 4) we rewrite (3.15) as

$$||w_2 - w_2'||_t - ||w_1 - w_1'||_t \leq 2\int_0^t ||w_2 - w_2'||_s \, ds$$

and note that the right hand side is nonnegative so the inequality remains true as

$$(||w_2 - w_2'||_t - ||w_1 - w_1'||_t)^+ \leq 2\int_0^t ||w_2 - w_2'||_s \, ds. \tag{3.16}$$

We now define

$$\begin{aligned} u_1(t) &= ||w_1 - w_1'||_t \,, \\ u_2(t) &= (||w_2 - w_2'||_t - ||w_1 - w_1'||_t)^+ \end{aligned}$$

and note

$$||w_2 - w_2'||_s \leq u_2(s) + u_1(s) \quad s \geq 0. \tag{3.17}$$

Then (3.14), (3.16), and (3.17) imply

$$\begin{aligned} u_1(t) &\leq 4\int_0^t u_1(s)ds + 2\int_0^t u_2(s)ds, \\ u_2(t) &\leq 2\int_0^t u_1(s)ds + 2\int_0^t u_2(s)ds. \end{aligned}$$

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 (3.17) 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 ξ. Suppose

$$(B^n, b^n, y^n) \to (B, b, y) \quad \text{as } n \to \infty.$$

Fix $\epsilon > 0$ and suppose w^n and w satisfy (3.8)-(3.9) 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 2 and the assumption $B^n \to B$. We have

$$\begin{aligned}
||w_1^n - w_1||_t &\leq |b^n - b| + ||y^n - y||_t \\
&\quad + \int_0^t ||-\phi_0(w_1^n) + \phi_0(w_1) + \phi_{B^n}(w_2^n) - \phi_B(w_2)||_s \, ds \\
&\leq \delta + \int_0^t \big(2\,||w_1^n - w_1||_s + ||\phi_{B^n}(w_2^n) - \phi_{B^n}(w_2)||_s \\
&\qquad + ||\phi_{B^n}(w_2) - \phi_B(w_2)||_s\big) ds \\
&\leq \delta + \int_0^t \left(2\,||w_1^n - w_1||_s + 2\,||w_2^n - w_2||_s + \delta\right) ds \\
&\leq \delta(1+t) + 2\int_0^t \left(||w_1^n - w_1||_s + ||w_2^n - w_2||_s\right) ds \qquad (3.18)
\end{aligned}$$

and

$$||w_2^n - w_2||_t \leq \delta(1+t) + ||w_1^n - w_1||_t + 2\int_0^t ||w_2^n - w_2||_s \, ds. \qquad (3.19)$$

As in the uniqueness argument above, we will apply Gronwall's inequality, with functions

$$u_1(t) = ||w_1^n - w_1||_t \quad \text{and} \quad u_2(t) = \left(||w_2^n - w_2||_t - ||w_1^n - w_1||_t\right)^+.$$

Then we have

$$\begin{aligned}
u_1(t) &\leq \delta(1+t) + 4\int_0^t u_1(s)ds + 2\int_0^t u_2(s)ds, \\
u_2(t) &\leq \delta(1+t) + 2\int_0^t u_1(s)ds + 2\int_0^t u_2(s)ds,
\end{aligned}$$

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

$$||w_1^n - w_1||_t \leq 2\delta(1+t)e^{6t} \quad \text{and} \quad ||w_2^n - w_2||_t \leq 4\delta(1+t)e^{6t}.$$

We choose $\delta = \frac{1}{4+4t}\epsilon e^{-6t}$ 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 9 imply that w_1 and w_2 are continuous if y_1 and y_2 are continuous. □

We are now prepared to prove Lemma 1.

Proof of Lemma 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 3.

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)-(2.8) imply that $z = (z_1, z_2)$ must be a solution of (3.8)-(3.9). By Lemma 3 this solution is unique. In particular, this solution is

$$\begin{aligned} x_1 &= f_1(b, y) = (\phi_0 \circ \xi_1)(b, y), \\ u_1 &= g_1(b, y) = (\psi_0 \circ \xi_1)(b, y), \\ x_2 &= f_2(b, y) = (\phi_B \circ \xi_2)(b, y), \\ u_2 &= g_2(b, y) = (\psi_B \circ \xi_2)(b, y). \end{aligned}$$

The reflection maps (ϕ_0, ψ_0) and (ϕ_B, ψ_B) are continuous in the uniform topology and also preserve continuity. Since ξ also has these properties by Lemma 3, we conclude that (f, g) are continuous and preserve continuity. □

3.2 Higher dimensions.

Because of the lack of reflection terms for $i \geq 3$, the higher dimensional terms in Theorem 1 behave in primarily one-dimensional ways. Therefore we will note an existence, uniqueness, and continuity result for a one-dimensional system that we will use in our proof of Theorem 1. This lemma is a special case of a slightly more general result proved by Pang, Talreja, and Whitt [14, Theorem 4.1], namely letting $h(x) = -x$.

Lemma 5. *Given $b \in \mathbb{R}$, and $y \in D$, consider*

$$x(t) = b + y(t) + \int_0^t -x(s)ds \tag{3.20}$$

$$x(t) \geq 0, \quad t \geq 0 \tag{3.21}$$

Then (3.20)-(3.21) *has a unique solution $x \in D$ so that there is a well defined function $f : \mathbb{R} \times D \to D$ mapping (b, y) into $x = f(b, y)$. Furthermore, the function (f, g) is continuous. Finally, if y is continuous, then so is x.*

With this lemma in hand, we can prove Theorem 1:

Proof of Theorem 1. We first prove existence, uniqueness, and preservation of continuity by induction on $k \geq i \geq 3$.

As the base case, observe that Lemma 5 implies there exists a unique solution $x_k(t)$ which preserves continuity of $y_k(t)$.

As an induction hypothesis, suppose for some $i \geq 3$ there exist unique solutions $x_{i+1}(t), \ldots, x_k(t)$ which are continuous if $y_{i+1}, \ldots, y_k$ are continuous, and consider $x_i(t)$.

We have the integral equation

$$\begin{aligned} x_i(t) &= b_i + y_i(t) + \int_0^t (-x_i(s) + x_{i+1}(s))ds \\ &= b_i + y_i(t) + \int_0^t x_{i+1}(s)ds + \int_0^t -x_i(s)ds. \end{aligned}$$

By the induction hypothesis $x_{i+1}(t)$ exists and is unique, so we let

$$\hat{y}(t) = y_i(t) + \int_0^t x_{i+1}(s)ds$$

and apply Lemma 5 with $y = \hat{y}$ to conclude that there exists a unique solution $x_i(t)$. This solution is continuous if $y_i, \ldots, y_k$ are continuous because in that case that $\hat{y}(t)$ is continuous.

By induction, we conclude that there exist unique solutions $x_3, \ldots, x_k$ which preserve continuity of $y_3, \ldots, y_k$. From this, we can define

$$\hat{y}_2(t) = y_2(t) + \int_0^t x_3(s)ds$$

and apply Lemma 1 with $y_2 = \hat{y}_2$ to complete the proof of existence, uniqueness, and preservation of continuity.

To verify that the map (f, g) is continuous, suppose $x^n(t)$ and $x(t)$ solve (2.4)-(2.8) for (B^n, b^n, y^n) and (B, b, y), respectively, and further suppose $(B^n, b^n, y^n) \to (B, b, y)$.

We again proceed by induction on $k \geq i \geq 3$.

For the base case note Lemma 5 implies $x_k^n \to x_k$.

As an induction hypothesis, suppose for some $i \geq 3$ we have $x_j^n \to x_j$ for $i \leq j \leq k$.

By the induction hypothesis we have

$$y_i^n(t) + \int_0^t x_{i+1}^n(s)ds \to y_i(t) + \int_0^t x_{i+1}(s)ds,$$

and thus, again by Lemma 5, $x_i^n \to x_i$.

By induction we conclude $x_i^n \to x_i$ for $3 \leq i \leq n$. This further implies

$$y_2^n(t) + \int_0^t x_3^n(s)ds \to y_2(t) + \int_0^t x_3(s)ds,$$

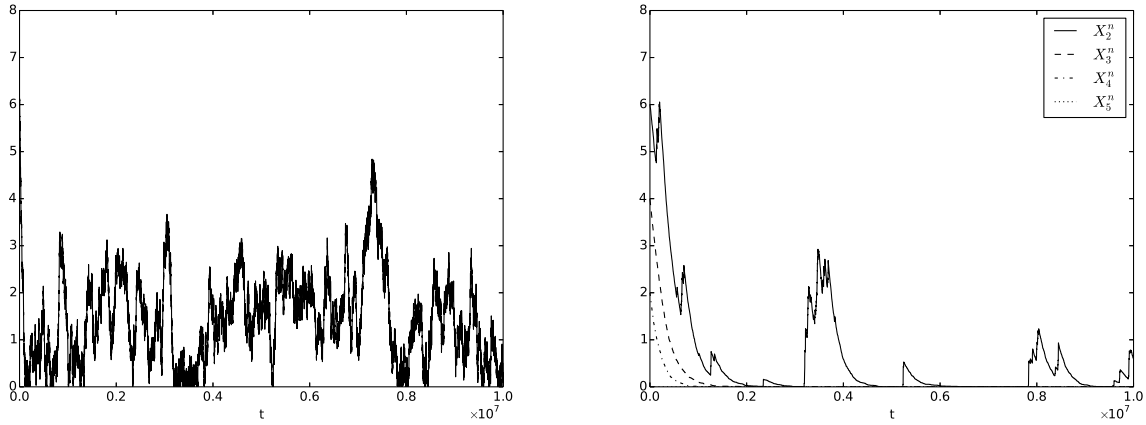
so Lemma 1 implies $(x_1^n, x_2^n) \to (x_1, x_2)$, and continuity is established. □

4 Truncation and martingale representation.

To use Theorem 1 in a CMT argument, we want to write the process X^n in the appropriate integral form. Instead of directly considering the full $M/M/n$-JSQ system, however, we will introduce a truncated variant which we will later show has the same behavior as X^n in the limit. It is this truncated system that will be shown to take the integral form of Theorem 1.

4.1 Truncation.

An important feature of the behavior of the $M/M/n$-JSQ system is that queues with more than one customer waiting are formed only if all queues have at least one customer waiting already. One of our goals is to show that Q_2^n, the number of queues with a customer waiting, is of the order $O(\sqrt{n})$ and thus it is unlikely for longer queues to form. With this in mind, as a proof technique, we now introduce a truncated version of the system in which no queues with length greater than 2 are created, though they are allowed to exist in the initial condition. This system will be significantly easier to analyze and will be shown to have stochastic behavior exactly matching the untruncated system with high probability. See Figure 1 to see a sample path of an original 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.



(a) Idle servers

(b) Longer queues

Figure 1: A simulated sample path of an $M/M/n$-JSQ system, showing the scaled number of idle servers (a) and queues of length at least two (X_2^n), three (X_3^n), four (X_4^n), and five (X_5^n). Simulated with $n = 10^5$, $\beta = 2.0$.

Now we make this more precise. Consider a system in which any arrival that would create a queue of length 3 or longer is rejected. That is, if an arrival occurs when all queues contain at least two customers, that arriving customer does not enter the system. We will denote this system $\hat{Q}^n = (\hat{Q}_1^n, \hat{Q}_2^n, \ldots)$. We also introduce scaled versions

$$\hat{X}_1^n(t) = \frac{\hat{Q}_1^n(t) - n}{\sqrt{n}} \quad \text{and} \quad \hat{X}_i^n(t) = \frac{\hat{Q}_i^n(t)}{\sqrt{n}}, \quad i \geq 2. \tag{4.1}$$

It will also be convenient for us to adopt the condition from Theorem 2 that $\hat{Q}_{k+1}^n(0) = 0$ for some $k \geq 2$. Since queues of length 3 or longer are never created in this system, this implies $\hat{Q}_i^n(t) = 0$ for $i \geq k+1$, and thus analysis of $\hat{Q}^n(t)$ can be limited to the first k dimensions.

We will now construct the truncated process $\hat{X}^n(t)$ 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.

4.2 Random time change.

We let A, D_i for $1 \leq i \leq k$ be rate-1 Poisson processes and write

$$\hat{Q}_1^n(t) = Q_1^n(0) + A\left(\lambda_n n t\right) - D_1\left(\int_0^t \left(\hat{Q}_1^n(s) - \hat{Q}_2^n(s)\right) ds\right) - \hat{U}_1^n(t), \tag{4.2}$$

$$\hat{Q}_2^n(t) = Q_2^n(0) + \hat{U}_1^n(t) - D_2\left(\int_0^t \left(\hat{Q}_2^n(s) - \hat{Q}_3^n(s)\right) ds\right) - \hat{U}_2^n(t), \tag{4.3}$$

$$\hat{Q}_i^n(t) = Q_i^n(0) - D_i\left(\int_0^t \left(\hat{Q}_i^n(s) - \hat{Q}_{i+1}^n(s)\right) ds\right), \tag{4.4}$$

$$\hat{Q}_k^n(t) = Q_k^n(0) - D_k\left(\int_0^t \hat{Q}_k^n(s) ds\right), \tag{4.5}$$

where $\hat{U}_1^n(t)$ is the number of arrivals in $[0,t]$ when every server has at least one customer, and $\hat{U}_2^n(t)$ is the number of arrivals in $[0,t]$ when every server has at least one customer *and* all n

servers have two customers. Formally, we define

$$\hat{U}_1^n(t) = \int_0^t \mathbb{1}\left\{\hat{Q}_1^n(s) = n\right\} dA(\lambda_n n s), \tag{4.6}$$

$$\hat{U}_2^n(t) = \int_0^t \mathbb{1}\left\{\hat{Q}_1^n(s) = n, \hat{Q}_2^n(s) = n\right\} dA(\lambda_n n s). \tag{4.7}$$

We can understand (4.2) term-by-term: first we record the initial state of the system with $Q_1^n(0)$, then arrivals are counted at their full rate $\lambda_n n$. The D_1 term represents departures, which occur as a Poisson process with rate equal to the number of customers in service. Since $\hat{Q}_1$ includes queues of length 1 and length 2, however, $\hat{Q}_1$ 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 $\hat{Q}_1^n(s) - \hat{Q}_2^n(s)$, the number of queues of length exactly 1 at time s. Through the first three terms of (4.2) we have recorded what the value of $\hat{Q}_1$ would be if it were not constrained to be at most n, so the final term will represent this barrier. The process $\hat{U}_1^n$ records any arrival which would increase $\hat{Q}_1$ above n, balancing the overcounting we get from $A(\lambda_n n t)$.

We can understand (4.3) 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 $\hat{U}_1^n$, so this will be the process we use to record potential increases to $\hat{Q}_2^n$. The process $\hat{U}_2^n$ provides the upper barrier n on $\hat{Q}_2^n$.

The remaining equations (4.4)-(4.5) are the same except that we record no arrivals, as our truncated approximation does not create queues of length 3 or longer.

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

4.3 Martingales.

Because our approach to (4.2)-(4.5) 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

$$\hat{M}_0^n(t) = \frac{1}{\sqrt{n}} A\left(\lambda_n n t\right) - \lambda_n \sqrt{n} t, \tag{4.8}$$

$$\hat{M}_i^n(t) = \frac{1}{\sqrt{n}} D_i\left(\int_0^t \left(\hat{Q}_i^n(s) - \hat{Q}_{i+1}^n(s)\right) ds\right) - \frac{1}{\sqrt{n}} \int_0^t \left(\hat{Q}_i^n(s) - \hat{Q}_{i+1}^n(s)\right) ds, \quad 1 \le i < k, \tag{4.9}$$

$$\hat{M}_k^n(t) = \frac{1}{\sqrt{n}} D_k\left(\int_0^t \hat{Q}_k^n(s) ds\right) - \frac{1}{\sqrt{n}} \int_0^t \hat{Q}_k^n(s) ds. \tag{4.10}$$

Via an argument exactly analogous to that of in §7.1 of [14] leading to Theorem 7.2 we obtain that $\hat{M}_i^n$ for $0 \le i \le k$ are square-integrable martingales with respect to an appropriate filtration. We note for later use that this argument also supplies the predictable quadratic variations

$$\left\langle \hat{M}_0^n \right\rangle (t) = \lambda_n t, \tag{4.11}$$

$$\left\langle \hat{M}_i^n \right\rangle (t) = \frac{1}{n} \int_0^t (\hat{Q}_i^n(s) - \hat{Q}_{i+1}^n(s)) ds, \tag{4.12}$$

$$\left\langle \hat{M}_k^n \right\rangle (t) = \frac{1}{n} \int_0^t \hat{Q}_k^n(s) ds. \tag{4.13}$$

We also define

$$\hat{V}_1^n(t) = \frac{\hat{U}_1^n(t)}{\sqrt{n}} \quad \text{and} \quad \hat{V}_2^n(t) = \frac{\hat{U}_2^n(t)}{\sqrt{n}}.$$

Then we have

$$\begin{aligned}
\hat{X}_1^n(t) &= \frac{\hat{Q}_1^n(t) - n}{\sqrt{n}} \\
&= \frac{Q_1^n(0) - n}{\sqrt{n}} + \frac{1}{\sqrt{n}} A\left(\lambda_n n t\right) \\
&\quad - \frac{1}{\sqrt{n}} D_1\left(\int_0^t (\hat{Q}_1^n(s) - \hat{Q}_2^n(s)) ds\right) - \frac{\hat{U}_1^n(t)}{\sqrt{n}} \\
&= X_1^n(0) + \hat{M}_0^n(t) + \lambda_n \sqrt{n} t - \hat{V}_1^n(t) \\
&\quad - \hat{M}_1^n(t) - \frac{1}{\sqrt{n}} \int_0^t \left(\hat{Q}_1^n(s) - \hat{Q}_2^n(s)\right) ds \\
&= X_1^n(0) + \hat{M}_0^n(t) - \hat{M}_1^n(t) + \lambda_n \sqrt{n} t - \hat{V}_1^n(t) \\
&\quad - \sqrt{n} t - \int_0^t \left(\frac{\hat{Q}_1^n(s) - n}{\sqrt{n}} - \frac{\hat{Q}_2^n(s)}{\sqrt{n}}\right) ds \\
&= X_1^n(0) + \hat{M}_0^n(t) - \hat{M}_1^n(t) - (1 - \lambda_n)\sqrt{n} t \qquad (4.14) \\
&\quad - \int_0^t (\hat{X}_1^n(s) - \hat{X}_2^n(s)) ds - \hat{V}_1^n(t),
\end{aligned}$$

and

$$\hat{X}_2^n(t) = X_2^n(0) + \hat{V}_1^n(t) - \hat{M}_2^n(t) - \int_0^t \left(\hat{X}_2^n(s) - \hat{X}_3^n(s)\right) ds - \hat{V}_2^n(t), \qquad (4.15)$$

$$\hat{X}_i^n(t) = X_i^n(0) - \hat{M}_i^n(t) - \int_0^t \left(\hat{X}_i^n(s) - \hat{X}_{i+1}^n(s)\right) ds, \quad 3 \le i \le k-1, \qquad (4.16)$$

$$\hat{X}_k^n(t) = X_k^n(0) - \hat{M}_k^n(t) - \int_0^t \hat{X}_k^n(s) ds. \qquad (4.17)$$

At this point we can also note that (4.14)-(4.17) put $\hat{X}^n(t)$ in the integral form of Theorem 1. The only difference is the processes $\hat{V}^n$, which are not described in exactly the same way. We see, however, that by (4.6) we have

$$\begin{aligned}
0 &= \int_0^\infty \mathbb{1}\{\hat{Q}_1^n(s) < n\} d\hat{U}_1^n(s) \\
&= \int_0^\infty \mathbb{1}\{\hat{X}_1^n(s) < 0\} d\hat{U}_1^n(s) \\
&= \int_0^\infty \mathbb{1}\{\hat{X}_1^n(s) < 0\} d\hat{V}_1^n(s). \qquad (4.18)
\end{aligned}$$

Similarly by (4.7) we have

$$0 = \int_0^\infty \mathbb{1}\{\hat{Q}_1^n(s) < n \text{ or } \hat{Q}_2^n(s) < n\} d\hat{U}_2^n(t).$$

which implies

$$0 = \int_0^\infty \mathbb{1}\{\hat{Q}_2^n(s) < n\} d\hat{U}_2^n(t)$$
$$= \int_0^\infty \mathbb{1}\{\hat{X}_2^n(s) < \sqrt{n}\} d\hat{V}_2^n(t). \tag{4.19}$$

Thus by (4.14)-(4.17) and (4.18)-(4.19) $\hat{X}^n$ is the unique solution of (2.4)-(2.8) for $b = X^n(0)$, $y_1 = \hat{M}_0^n(t) - \hat{M}_1^n(t) - (1-\lambda_n)\sqrt{n}t$, $y_i = -\hat{M}_i^n(t)$, $2 \le i \le k$, and $B = \sqrt{n}$. Thus to apply the CMT it remains to prove the convergence of the martingales (4.8)-(4.10).

5 Martingale convergence.

We will now prove the convergence of $\hat{M}_i^n$ to Brownian motions.

Lemma 6. *For the sequences of scaled martingales $\hat{M}_i^n$ defined in Section 4.3 we have the convergence*

$$\left(\hat{M}_0^n, \hat{M}_1^n, \hat{M}_2^n, \ldots, \hat{M}_k^n\right) \Rightarrow (W_1, W_2, 0, \ldots, 0) \quad \text{in } D \quad \text{as } n \to \infty, \tag{5.1}$$

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 restate here convenience.

Lemma 7. *(FCLT for independent Poisson processes) If A and D_i are independent rate-1 Poisson processes and*

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

for $C = A, D_i$, then

$$(M_{A,n}, M_{D_1,n}, \ldots, M_{D_k,n}) \Rightarrow (W_1, W_2, \ldots, W_{k+1}) \quad \text{in } D^{k+1} \quad \text{as } n \to \infty$$

where W_i are independent standard Brownian motions.

To apply Lemma 7 we will define random and deterministic time changes such that the martingales $\hat{M}_i^n$ can be written as a composition of a time change and the scaled Poisson processes $M_{C,n}$. Specifically, let

$$\Phi_{A,n}(t) = \lambda_n t, \tag{5.2}$$
$$\Phi_{D_i,n}(t) = \frac{1}{n}\int_0^t \hat{Q}_i^n(s)ds - \frac{1}{n}\int_0^t \hat{Q}_{i+1}^n(s)ds, \tag{5.3}$$
$$\Phi_{D_k,n}(t) = \frac{1}{n}\int_0^t \hat{Q}_k^n(s)ds, \tag{5.4}$$

so that we have

$$\hat{M}_0^n = \hat{M}_{A,n} \circ \Phi_{A,n}, \quad \hat{M}_i^n = \hat{M}_{D_i,n} \circ \Phi_{D_i,n}, \quad 1 \le i \le k.$$

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

$$\Phi_{A,n} \Rightarrow e \quad \text{as } n \to \infty, \tag{5.5}$$

where e is the identity function in D.

Next we note that the terms of (5.3) interleave over i, so for $1 \le i \le k-1$ we have

$$\Phi_{D_i,n} \Rightarrow f_i - f_{i+1} \text{ as } n \to \infty,$$

where f_i is the limit of $\tilde{\Phi}_{D_i,n}$ with

$$\tilde{\Phi}_{D_i,n}(t) = \frac{1}{n}\int_0^t \hat{Q}_i^n(s)ds.$$

To find f_i we will first show fluid limits for $\hat{Q}_i^n$.

Lemma 8. *Let Ψ_i^n for $1 \le i \le k$ be defined by*

$$\Psi_i^n(t) = \frac{\hat{Q}_i^n(t)}{n}, \quad t \ge 0.$$

Then for $2 \le i \le k$,

$$\Psi_1^n \Rightarrow \omega \quad \text{and} \quad \Psi_i^n \Rightarrow 0 \quad \text{as } n \to \infty \tag{5.6}$$

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

The proof of this lemma is found in Section 5.1. To use Lemma 8 we define a continuous function $h : D \to D$ by

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

for $t \ge 0$. The $\tilde{\Phi}_{D_1,n} = h \circ \Psi_n$ so by the CMT and Lemma 8 we know $f_1 = h \circ \omega$. Namely,

$$f_1(t) = \int_0^t 1ds = t$$

so $f_1 = e$ is the identity function in D. Therefore we have

$$\tilde{\Phi}_{D_1,n} \Rightarrow e \quad \text{as } n \to \infty.$$

For $2 \le i \le k$ we have $f_i(t) = \int_0^t 0ds = 0$ so $f_i = 0$ on D. We conclude that

$$\Phi_{D_1,n} \Rightarrow e \quad \text{as } n \to \infty, \tag{5.7}$$

and for $2 \le i \le k$

$$\Phi_{D_i,n} \Rightarrow 0 \quad \text{as } n \to \infty. \tag{5.8}$$

Therefore once we establish Lemma 8 we can prove Lemma 6:

Proof of Lemma 6. We apply the CMT with Lemma 7 and the limits (5.5), (5.7), and (5.8) to obtain

$$\begin{aligned}\left(\hat{M}_0^n, \hat{M}_1^n, \hat{M}_2^n, \ldots, \hat{M}_k^n\right) &= (M_{A,n} \circ \Phi_{A,n}, M_{D_1,n} \circ \Phi_{D_1,n}, M_{D_2,n} \circ \Phi_{D_2,n}, \ldots, M_{D_k,n} \circ \Phi_{D_k,n}) \\ &\Rightarrow (W_1 \circ e, W_2 \circ e, W_3 \circ 0, \ldots, W_{k+1} \circ 0) \\ &= (W_1, W_2, 0, \ldots, 0).\end{aligned}$$

□

5.1 Fluid limit.

We will prove Lemma 8 by showing that $\hat{X}^n$ is stochastically bounded in D. Namely, we will prove that the sequence of real-valued random variables $\left|\left|\hat{X}_i^n\right|\right|_t$ is tight for every $t > 0$ and $i \geq 1$. For a more complete discussion of stochastic boundedness as we use it here see §5 of [14].

The stochastic boundedness of $\hat{X}_1^n$ and $\hat{X}_2^n$ will follow from the stochastic boundedness of $\hat{M}_0^n$, $\hat{M}_1^n$, $\hat{M}_2^n$, and $\hat{X}_3^n$. To see this, we prove

Lemma 9. *Given $(B_n, X_i^n(0), Y_i^n)$ a random element of $\bar{\mathbb{R}}_+ \times \mathbb{R} \times D$ for each $n \geq 1$ and $i = 1, 2$, recall that Lemma 1 implies that the system*

$$\begin{aligned}
\hat{X}_1^n(t) &= X_1^n(0) + Y_1^n(t) + \int_0^t (-\hat{X}_1^n(s) + \hat{X}_2^n(s))ds - \hat{V}_1^n(t), \\
\hat{X}_2^n(t) &= X_2^n(0) + Y_2^n(t) + \int_0^t (-\hat{X}_2^n(s))ds + \hat{V}_1^n(t) - \hat{V}_2^n(t) \geq 0, \\
0 &= \int_0^\infty \mathbb{1}\{\hat{X}_1^n(t) < 0\} d\hat{V}_1^n(t), \\
0 &= \int_0^\infty \mathbb{1}\{\hat{X}_2^n(t) < B_n\} d\hat{V}_2^n(t),
\end{aligned}$$

has a unique solution $(\hat{X}^n, \hat{V}^n)$. If the sequences $(X^n(0), n \geq 1)$ and $(Y_i^n, n \geq 1)$ are stochastically bounded for $i = 1, 2$, then the sequence $(\hat{X}^n, n \geq 1)$ 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

$$\left|\left|\hat{X}^n\right|\right|_t \leq 8e^{6t} \left(|X^n(0)| + ||Y^n||_t\right), \tag{5.9}$$

from which the result follows.

To show (5.9), we will prove a similar bound for the unreflected process W^n defined by Lemma 3. Then (5.9) will follow from the Lipschitz continuity of the reflection maps ϕ_0 and ϕ_{B_n}.

Just as in Lemma 1 and Lemma 3 we write $\hat{X}_1^n(t) = \phi_0(W_1^n(t))$ and $\hat{X}_2^n(t) = \phi_{B^n}(W_2^n(t))$ where $W_1^n(t)$ and $W_2^n(t)$ satisfy

$$W_1^n(t) = X_1^n(0) + Y_1^n(t) + \int_0^t (-\phi_0(W_1^n(s)) + \phi_{B^n}(W_2^n(s)))ds, \tag{5.10}$$

$$W_2^n(t) = X_2^n(0) + Y_2^n(t) + \int_0^t (-\phi_{B^n}(W_2^n(s)))ds + \psi_0\left(W_1^n(t)\right). \tag{5.11}$$

We now use Gronwall's inequality as stated in Lemma 4. Using the Lipschitz property for ϕ_0,ϕ_{B_n}, and ψ_0 we have for $t \geq 0$

$$\begin{aligned}
||W_1^n||_t &\leq |X_1^n(0)| + ||Y_1^n||_t + 2\int_0^t \left(||W_2^n||_s + ||W_1^n||_s\right) ds, \\
||W_2^n||_t &\leq |X_2^n(0)| + ||Y_2^n||_t + ||\psi_0(W_1^n)||_t + \int_0^t ||W_2^n||_s \, ds.
\end{aligned}$$

Now we note that we have

$$||\psi_0(W_1^n)||_t \leq ||W_1^n||_t .$$

We define

$$u_1(t) = ||W_1^n||_t \quad \text{and} \quad u_2(t) = \left(||W_2^n||_t - ||W_1^n||_t\right)^+.$$

Finally we note

$$||W_2^n||_t \leq u_2(t) + u_1(t), \tag{5.12}$$

so we can write the inequalities

$$u_1(t) \leq |X_1^n(0)| + ||Y_1^n||_t + 4\int_0^t u_1(s)ds + 2\int_0^t u_2(s)ds,$$
$$u_2(t) \leq |X_2^n(0)| + ||Y_2^n||_t + \int_0^t u_1(s)ds + \int_0^t u_2(s)ds.$$

Let $|X^n(0)| + ||Y^n||_t = K$. Then Lemma 4 implies

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

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

$$||W_1^n||_t \leq 2Ke^{6t} \quad \text{and} \quad ||W_2^n||_t \leq 4Ke^{6t}.$$

Since ϕ_0 and ϕ_{B^n} are Lipschitz continuous with constant 2 this implies

$$\left|\left|\hat{X}_1^n\right|\right|_t \leq 4Ke^{6t} \quad \text{and} \quad \left|\left|\hat{X}_2^n\right|\right|_t \leq 8Ke^{6t},$$

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 3, and that it does not use any of the continuity properties proved using that boundedness.

In our application of Lemma 9 we will have $Y_1^n = \hat{M}_1^{n,1} - \hat{M}_1^{n,2} - (1-\lambda_n)\sqrt{n}t$ and $Y_2^n(t) = \hat{M}_2^{n,2}(t) + \int_0^t \hat{X}_3^n(s)ds$, so it remains to prove that $\int_0^t \hat{X}_3^n(s)ds$ and each martingale $\hat{M}_k^{n,i}$ is stochastically bounded.

To show that $\int_0^t \hat{X}_3^n(s)ds$ is stochastically bounded we need only show that $\hat{X}_3^n$ is stochastically bounded. This follows from the fact that $Q_3^n(0)$ is stochastically bounded by assumption (2.9) and

$$\hat{X}_3^n(t) = \frac{\hat{Q}_3^n(t)}{\sqrt{n}} \leq \frac{Q_3^n(0)}{\sqrt{n}}$$

because no queues of length 3 or longer are ever created. An identical argument proves stochastic boundedness of $\hat{X}_i^n$ for $4 \leq i \leq k$.

To prove the stochastic boundedness of these martingales we will use the following lemma from [14]:

Lemma 10 ([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 $\hat{M}_i^n$ are stochastically bounded. In the case of $\hat{M}_0^n$ this is immediate since by (4.11) the quadratic variation is deterministic.

For $\hat{M}_1^n$ we refer to (4.12) and apply crude bounds to see

$$\begin{aligned}\left\langle \hat{M}_1^n \right\rangle (t) &= \frac{1}{n}\int_0^t (\hat{Q}_1^n(s) - \hat{Q}_2^n(s))ds \\ &\leq \frac{1}{n}\int_0^t \hat{Q}_1^n(s)ds \\ &\leq \frac{t}{n}\left(Q_1^n(0) + A(\lambda_n nt)\right).\end{aligned}$$

It suffices to show stochastic boundedness of each term in the sum. For $Q_1^n(0)$ this follows from assumption (2.9).

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

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

with probability 1, which implies stochastic boundedness, so we conclude that $\hat{M}_1^n$ is stochastically bounded.

For $\hat{M}_2^n$ we have

$$\begin{aligned}\left\langle \hat{M}_2^n \right\rangle (t) &\leq \frac{t}{n}\left(Q_2^n(0) + \hat{U}_1^n(t)\right) \\ &\leq \frac{t}{n}\left(Q_2^n(0) + A(\lambda_n nt)\right),\end{aligned}$$

and stochastic boundedness follows.

We now return to the proof of Lemma 8:

Proof of Lemma 8. We have for $i \leq 2 \leq k$ that

$$\hat{X}_1^n = \frac{\hat{Q}_1^n - n}{\sqrt{n}} \quad \text{and} \quad \hat{X}_i^n = \frac{\hat{Q}_i^n}{\sqrt{n}}$$

are stochastically bounded. Therefore

$$\frac{\hat{X}_i^n}{\sqrt{n}} \Rightarrow 0 \quad \text{in } D \quad \text{as } n \to \infty.$$

From the definition of $\hat{X}^n$ this is equivalent to

$$\Psi_1^n = \frac{\hat{Q}_1^n}{n} \Rightarrow \omega \quad \text{and} \quad \Psi_i^n = \frac{\hat{Q}_i^n}{n} \Rightarrow 0 \quad \text{in } D \quad \text{as } n \to \infty$$

for $2 \leq i \leq k$. □

5.2 Proof of Theorem 2.

Now that we have the convergence of the martingale processes $\hat{M}_i^n$, we can apply the CMT to prove Theorem 2.

Proof of Theorem 2. We first show $\hat{X}^n \Rightarrow X$.

In Theorem 1, in the pre-limit regime we set $B_n = \sqrt{n}$, $b_i = X_i^n(0)$ for $1 \leq i \leq k$,

$$y_1(t) = \hat{M}_0^n(t) - \hat{M}_1^n(t) - (1-\lambda_n)\sqrt{n}t,$$

and $y_i(t) = -\hat{M}_i^n(t)$ for $2 \le i \le k$. Equations (4.18) and (4.19) show that $\hat{V}_1^n$ and $\hat{V}_2^n$ are appropriately acting as u_1 and u_2 in the integral representation, so $x_i(t) = \hat{X}_i^n(t)$ for $1 \le i \le k$. For application of the CMT we need only determine the limits of B, b and y. We have $B_n \to \infty$, so in the limit $u_2 = 0$.

By assumption we have

$$\hat{X}_k^n(0) \Rightarrow X_k(0),$$

so in the limiting system we let $b_i = \hat{X}_i(0)$. Next we have by (2.1) and (5.1)

$$\begin{aligned}\hat{M}_0^n(t) - \hat{M}_1^n(t) - (1-\lambda_n)\sqrt{n}t &\Rightarrow W_1(t) - W_2(t) - \beta t \\ &\stackrel{d}{=} \sqrt{2}W(t) - \beta t,\end{aligned}$$

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

$$-\hat{M}_i^n \Rightarrow 0$$

for $2 \le i \le k$ so in the limiting system we have $y_1(t) = \sqrt{2}W(t) - \beta t$ and $y_i(t) = 0$, for $2 \le i \le k$.

The CMT then implies that for $1 \le i \le k$, $\hat{X}_i^n \Rightarrow X_i$ in D as $n \to \infty$ where X_i is described by (2.10)-(2.15). We can augment the truncated system with $\hat{X}_i^n(t) = 0$ for $i > k$ and note that $0 = \hat{X}_i^n \Rightarrow X_i = 0$ for such i. Therefore $\hat{X}^n \Rightarrow X$.

Now we consider the untruncated system described by X^n. By an argument like that Section 4, we have

$$Q_1^n(t) = Q_1^n(0) + A\left(\lambda_n n t\right) - D_1\left(\int_0^t \left(Q_1^n(s) - Q_2^n(s)\right) ds\right) - U_1^n(t), \tag{5.13}$$

$$Q_i^n(t) = Q_i^n(0) + U_{i-1}^n(t) - D_i\left(\int_0^t \left(Q_i^n(s) - Q_{i+1}^n(s)\right) ds\right) - U_i^n(t), \quad i \ge 2, \tag{5.14}$$

where $U_i^n(t)$ is the number of arrivals in $[0,t]$ when every server has at least i customers. Note that we introduce extra rate one Poisson processes D_i for $i > k$.

Now define

$$t_n^* = \inf\{t \ge 0 : Q_2^n(t) = n\} \tag{5.15}$$

and note that for $t \in [0, t_n^*)$ we have $U_i^n(t) = 0$ for $i \ge 2$. This, along with $Q_i^n(0) = 0$ for $i > k$, implies that for such t, the system (5.13)-(5.14) becomes

$$\begin{aligned}
Q_1^n(t) &= Q_1^n(0) + A\left(\lambda_n n t\right) - D_1\left(\int_0^t \left(Q_1^n(s) - Q_2^n(s)\right) ds\right) - U_1^n(t), \\
Q_2^n(t) &= Q_2^n(0) + U_1^n(t) - D_2\left(\int_0^t \left(Q_2^n(s) - Q_3^n(s)\right) ds\right) - U_2^n(t), \\
Q_i^n(t) &= Q_i^n(0) - D_i\left(\int_0^t \left(Q_i^n(s) - Q_{i+1}^n(s)\right) ds\right), \\
Q_k^n(t) &= Q_k^n(0) - D_k\left(\int_0^t Q_k^n(s) ds\right),
\end{aligned}$$

which precisely matches (4.2)-(4.5). Thus for $t \in [0, t_n^*)$, $X^n(t)$ and $\hat{X}^n(t)$ are identical.

It only remains to show for all $t \ge 0$ that $\mathbb{P}(t_n^* \le t) \to 0$ as $n \to \infty$. Because the systems are identical up to time t_n^*, we can replace $Q_2^n(t)$ in (5.15) with $\hat{Q}_2^n(t)$ to see

$$\begin{aligned}\mathbb{P}(t_n^* \le t) &= \mathbb{P}\left(\sup_{0 \le s \le t} \hat{Q}_2^n(s) \ge n\right) = \mathbb{P}\left(\sup_{0 \le s \le t} \hat{X}_2^n(s) \ge \sqrt{n}\right) \\ &\le \mathbb{P}\left(\sup_{0 \le s \le t} \hat{X}_2^n(s) \ge C\right)\end{aligned}$$

for constant $0 < C \leq \sqrt{n}$. By the weak convergence $\hat{X}_2^n \Rightarrow X_2$ and the fact that $\left\{\sup_{0\leq s\leq t} \hat{X}_2^n(s) \geq C\right\}$ is closed, we have

$$\limsup_n \mathbb{P}\left(\sup_{0\leq s\leq t} \hat{X}_2^n(s) \geq C\right) \leq \mathbb{P}\left(\sup_{0\leq s\leq t} X_2(s) \geq C\right)$$

By continuity of probability we have

$$\lim_{C\to\infty} \mathbb{P}\left(\sup_{0\leq s\leq t} X_2(s) \geq C\right) = \mathbb{P}\left(\sup_{0\leq s\leq t} X_2(s) = \infty\right) = 0.$$

We therefore have

$$\limsup_n \mathbb{P}(t_n^* \leq t) \leq \lim_{C\to\infty} \limsup_n \mathbb{P}\left(\sup_{0\leq s\leq t} \hat{X}_2^n(s) \geq C\right) \leq \lim_{C\to\infty} \mathbb{P}\left(\sup_{0\leq s\leq t} X_2(s) \geq C\right) = 0$$

and thus $\mathbb{P}(t_n^* \leq t) \to 0$ as $n \to \infty$. We conclude $X^n \Rightarrow X$. □

6 Waiting time.

An important performance measure of a queueing system is the expected time that customers will have to wait before entering service. In the $M/M/n$ system with a single queue in the Halfin-Whitt regime, the expected waiting time is of the order $O\left(1/\sqrt{n}\right)$. We will now show that the $M/M/n$-JSQ system has the same order of aggregate waiting time in the transient regime, and thus seems to have a minimal loss of efficiency as measured by waiting time.

Notice that our representation of the system allows us to directly consider the total time any customers in the system will wait. In particular, the instantaneous number of customers waiting to be served at a given time t is precisely $\sum_{i\geq 2} Q_i^n(t)$. This quantity can be integrated over time to compute the aggregate waiting time in the system. With this insight, we prove the following:

Theorem 3. *The aggregate waiting time of customers who arrived over the time period $[0,t]$, which we denote Z_t^n, satisfies*

$$\lim_{C\to\infty} \limsup_n \mathbb{P}(Z_t^n \geq C\sqrt{n}) = 0. \tag{6.1}$$

Since the total number of arrivals to the system in this time is $\Theta(n)$ with high probability, the waiting time per arrival is $O\left(1/\sqrt{n}\right)$ with high probability.

Proof. As noted above, the aggregate waiting time over the period $[0,t]$ is

$$Z_t^n = \int_0^t \sum_{i\geq 2} Q_i^n(s) ds.$$

We consider a scaled version $Y_t^n = Z_t^n/\sqrt{n}$. Note that $Y_t^n \leq t \sup_{0\leq s\leq t} \sum_{i\geq 2} X_i^n(s)$, and thus

$$\mathbb{P}\left(Y_t^n \geq C\right) \leq \mathbb{P}\left(t \sup_{0\leq s\leq t} \sum_{i\geq 2} X_i^n(s) \geq C\right)$$

for any constant $C > 0$. By the weak convergence $X_i^n \Rightarrow X_i$ and the fact that $\{t \sup_{0\leq s\leq t} \sum_{i\geq 2} X_i^n(s) \geq C\}$ is closed, we have

$$\limsup_n \mathbb{P}\left(t \sup_{0\leq s\leq t} \sum_{i\geq 2} X_i^n(s) \geq C\right) \leq \mathbb{P}\left(t \sup_{0\leq s\leq t} \sum_{i\geq 2} X_i(s) \geq C\right)$$

By continuity of probability and $X_i = 0$ for $i > k$ we have

$$\lim_{C\to\infty} \mathbb{P}\left(t \sup_{0\le s\le t} \sum_{i\ge 2} X_i(s) \ge C \right) = 0.$$

Therefore we conclude

$$\lim_{C\to\infty} \limsup_n \mathbb{P}\left(Y_t^n \ge C\right) \le \lim_{C\to\infty} \mathbb{P}\left(t \sup_{0\le s\le t} \sum_{i\ge 2} X_i(s) \ge C \right) = 0,$$

which proves (6.1). We note that this implies $Y_t^n = O(1)$ with high probability, and thus $Z_t^n = O(\sqrt{n})$ with high probability.

Because customers arrive according to a Poisson process with rate $\lambda_n n = \Theta(n)$, this implies the waiting time per arrival is $O(\sqrt{n}/n) = O(1/\sqrt{n})$ with high probability, completing the proof. □

Theorem 3 does not directly tell us anything about the distribution of the waiting time. We can see from considering the system that this waiting time is distributed in a qualitatively different way than the standard $M/M/n$ system. 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. Because the aggregate waiting time is of the order $O(\sqrt{n})$ and and any arriving customers who wait at all incur a constant order waiting time, the total number of customers who have to wait is also $O(\sqrt{n})$. As noted above, the total number of arriving customers is order n, so the fraction of customers who have to wait is of the order $O(1/\sqrt{n})$.

7 Open questions.

Theorem 2 proves that the behavior of the $M/M/n$-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 $O(\sqrt{n})$. If long queues are initially present, they will empty in fixed time, and no additional long queues will be created.

Significant questions 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).

Finally it is always of interest to analyze our system for general interarrival and, especially, general service times distribution. We conjecture that the qualitative behavior 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] Hanqin 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.