Operations Research Letters 40 (2012) 307–312

Contents lists available at SciVerse ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading

Yunan Liu, Ward Whitt *

Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695, USA
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA

ARTICLE INFO

Article history:
Received 25 October 2011
Accepted 25 May 2012
Available online 4 June 2012

Keywords:
Many-server heavy-traffic limit
Functional weak law of large numbers
Queues with time-varying arrivals
Nonstationary queues
Deterministic fluid model
Non-Markovian queues

ABSTRACT

A many-server heavy-traffic functional weak law of large numbers is established for the $G_t/GI/s_t + GI$ queueing model, which has customer abandonment (the $+GI$), time-varying arrival rate and staffing (the subscript t) and non-exponential service and patience distributions (the two GI's). This limit provides support for a previously proposed deterministic fluid approximation, and extends a previously established limit for the special case of exponential service times.

1. Introduction

This paper is a sequel to our two previous papers [9,10]. In [9] we introduced and analyzed a deterministic fluid model that serves as an approximation for the many-server $G_t/GI/s_t + GI$ queueing model, which has customer abandonment (the $+GI$), time-varying arrival rate and staffing (the subscript t), unlimited waiting space, the first-come first-served service discipline and non-exponential service and patience distributions (the two GI's).

The fluid model in [9] is a time-varying extension of the $G/GI/s + GI$ fluid model introduced in [19]. In [9], the system was assumed to alternate between overloaded (OL) and underloaded (UL) intervals. In order to achieve greater mathematical tractability, the system was assumed to be critically loaded only at isolated switching points between the OL and UL intervals. Time-varying arrival rates commonly occur in service systems [4]. The alternating behavior commonly occurs when it is difficult to dynamically adjust the staffing level in response to changes in demand. If the staffing cannot be changed rapidly enough, then system managers must choose fixed or nearly fixed staffing levels that respond to several levels of demand over a time interval. Then it may not be cost-effective to staff at a consistently high level in order to avoid overloading at any time. Then the fluid model may capture the essential performance.

Extensive simulation experiments with numerical examples, such as the one reported in Section 2 of [9], confirm that the performance functions of the deterministic $G_t/GI/s_t + GI$ fluid model in [9] can readily be computed and that they coincide with the many-server heavy-traffic limit of the corresponding sequence of scaled stochastic processes in the $G_t/GI/s_t + GI$ stochastic queueing model. Our more recent paper [10] provides mathematical justification by establishing many-server heavy-traffic limits. Those limits provide mathematical support for both the previous fluid approximation and a refined Gaussian process approximation. However, those results were restricted to the special case of exponential (M) service times. The purpose of the present paper is to complete the mathematical justification of the fluid model by extending the fluid limit portion of those asymptotic results to the case of GI service.

As in [10], for the results here we draw on associated limits for two-parameter stochastic processes arising in infinite-server models in [14], which in turn draws on [8]. Our limits for many-server queues with time varying arrival rate and staffing extend earlier limits for the Markovian $M_t/M/s_t + M$ model in [12] and a discrete-time limit for the $G_t/GI/s + GI$ model in [19]. Related limits for $G/GI/s + GI$ models are contained in [5–7,15,16,21] and references therein.

Here is how this paper is organized. In Section 2 we state the main new result and in Section 3 we outline the proof. In Section 4 we state and prove Theorem 2, establishing a FWLLN for the two-parameter process representing the number of customers that were initially in service at time 0 and remain in service at time t, and have been so far a duration at most y (including a

* Correspondence to: Mail Code 4704, S. W. Mudd Building, 500 West 120th Street, New York, NY 10027-6699, USA.
E-mail address: ww2040@columbia.edu (W. Whitt).

0167-6377/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2012.05.010

period prior to time 0), as a function of t and y. Theorem 2 is an important complement to existing results for infinite-server models, extending [8,14], where it is assumed that the remaining service times of customers initially in service are i.i.d.. In Sections 5 and 6, respectively, we provide extra details about the proof for UL and OL intervals. Finally, in Section 7 we provide a longer proof of one lemma used in the proof of Theorem 2.

2. The new FWLLN

We consider a sequence of $G_t/GI/s_t + GI$ queueing models indexed by n. Model n has a general arrival process with time-varying arrival rate $\lambda_n(t) \equiv n\lambda(t)$ (with $\equiv$ denoting equality by definition), i.i.d. service times with cumulative distribution function (cdf) G, a time-varying number of servers $s_n(t) \equiv \lceil ns(t) \rceil$ (the least integer above $ns(t)$) and customer abandonment from queue, where the patience times of successive customers to enter the queue are i.i.d. with general cdf F. The two cdf's G and F are fixed independent of n, and differentiable, with positive probability density functions (pdf's) g and f. Our scaling of the fixed functions λ and s induces the familiar many-server heavy-traffic scaling; the functions λ and s are the arrival rate and staffing level in the associated fluid model, assumed to be bounded away from zero. The arrival process, service times and patience times are mutually independent. New arrivals enter service immediately if there is a free server; otherwise they join the queue, from which they enter service in order of arrival, if they do not first abandon.

Let $B_n(t, y)$ ($Q_n(t, y)$) denote the number of customers in service (queue) at time t that have been so for time at most y. Let $B_n(t) \equiv B_n(t, \infty)$ ($Q_n(t) \equiv Q_n(t, \infty)$), the total number of customers in service (queue). Let $X_n(t) \equiv B_n(t) + Q_n(t)$, the total number of customers in the system. Let $W_n(t)$ be the head-of-line waiting time (HWT), i.e., the elapsed waiting time for the customer at the head of the line at time t (the customer who has been waiting the longest). Let $V_n(t)$ be the potential waiting time (PWT) at time t, i.e., the virtual waiting time at time t (the waiting time if there were a new arrival at time t) assuming that that customer never would abandon (but without actually altering any arrival's abandonment behavior). Let $A_n(t)$ be the number of abandonments and let $D_n(t)$ be the number of departures (service completions) in the interval $[0, t]$.

Let the associated FWLLN-scaled or fluid-scaled processes be

$$\begin{aligned} &\bar{B}_n(t, y) \equiv n^{-1}B_n(t, y), \qquad \bar{Q}_n(t, y) \equiv n^{-1}Q_n(t, y), \\ &\bar{X}_n(t) \equiv n^{-1}X_n(t), \qquad \bar{D}_n(t) \equiv n^{-1}D_n(t), \\ &\bar{A}_n(t) \equiv n^{-1}A_n(t). \end{aligned} \tag{1}$$

The waiting times $W_n(t)$ and $V_n(t)$ are not scaled in the fluid limit. The same notation without the subscript n and without the bar is used to denote the deterministic limits.

The limits are established in the function space $\mathbb{D} \equiv \mathbb{D}([0, T], \mathbb{R})$ with the usual Skorohod J_1 topology and metric d_{J_1} [2,17,20] and products of that space with the product topology. Since all limits will be continuous functions, convergence is equivalent to uniform convergence over the compact time interval. For the two-parameter processes, the limits hold in the space $\mathbb{D}_{\mathbb{D}} \equiv \mathbb{D}([0, T], \mathbb{D}([0, T], \mathbb{R}))$ of $\mathbb{D}$-valued functions. Since the space $(\mathbb{D}, J_1)$ is a complete separable metric space, this space of $\mathbb{D}$-valued functions falls within Skorohod's [17] original framework. See [14,18] for additional details.

We make all assumptions in [9,10], allowing GI service with assumptions in [9] instead of requiring the exponential (M) service as in [10]; see [9,10] for full details. First, Assumption 7 of [9] stipulates that the fluid model has only finitely many switches between OL and UL intervals in any finite time interval. Sufficient conditions are given in [11]. The system is assumed to start in a UL interval. Moreover, the staffing function is assumed to be feasible for the limiting fluid model. (No fluid in service is forced out by staffing reductions. In Section 9 of [9] we give an algorithm to find the minimum feasible staffing function greater than or equal to any given one.) FWLLN's are assumed to hold for the arrival counting process, with deterministic limit $\Lambda(t) \equiv \int_0^t \lambda(u)\,du$, and for the initial content in service: $\bar{B}_n(0, \cdot) \Rightarrow B(0, \cdot)$ in $\mathbb{D}$. Assumption 2 of [9] requires that the functions s, Λ, G, F and $B(0, \cdot)$ be differentiable with piecewise-continuous derivatives $\dot{s}$, λ, g, f and $b(0, \cdot)$ in $\mathbb{D}$.

Assumption 8 of [9] requires a bound on the tail of the initial fluid density in service with respect to the service time pdf: $\tau^{\uparrow}(b, g, T) < \infty$ for each $T > 0$, where

$$\tau^{\uparrow}(b, g, T) \equiv \sup_{0 \le s \le T} \tau(b, g, s),$$

$$\tau(b, g, s) \equiv \int_0^{\infty} \frac{b(0, x)g(s + x)}{G^c(x)}\,dx \tag{2}$$

with $G^c(x) \equiv 1 - G(x) > 0$ and $b(t, x)$ the fluid content density in service at time t that has been in service for a duration x.

In [9] we gave several convenient sufficient conditions to have $\tau(b, g, T) < \infty$ for each $T > 0$. One is $B(0) - B(0, y^{\uparrow}) = 0$ for some $y^{\uparrow} > 0$.

Assumption 1 (*Bound On Time Initial Content Has Been in Service*). There exist $y^{\uparrow} > 0$ such that $B_n(0) - B_n(0, y^{\uparrow}) = 0$ for all $n \ge 1$. Moreover, there exists $x^{\uparrow}$ such that $b(0, x) > 0, 0 < x < x^{\uparrow}$ and $b(0, x) = 0, x > x^{\uparrow}$.

The first part of Assumption 1 holds if the systems start empty some time in the finite past.

Theorem 1 (*FWLLN*). *If, in addition to the assumptions of* [9,10], *Assumption 1 holds, then the FWLLN established in* [10] *for the $G_t/M/s_t+GI$ model holds for the corresponding $G_t/GI/s_t+GI$ model, i.e.,*

$$(\bar{X}_n, \bar{D}_n, \bar{Q}_n, \bar{B}_n, \bar{A}_n, W_n, V_n) \Rightarrow (X, D, Q, B, A, W, V) \quad \text{as } n \to \infty \tag{3}$$

in $\mathbb{D}^2 \times \mathbb{D}_{\mathbb{D}}^2 \times \mathbb{D}^3$, where the converging processes are defined in (1) and the limit (X, D, Q, B, A, W, V) is the vector of continuous deterministic functions defined and characterized in [9], *having $Q(t) \ge 0$, $X(t) = Q(t) + s(t)$ and $B(t) = s(t)$ for all t during each OL interval and $Q(t) = 0$, $X(t) = B(t)$ and $B(t) \le s(t)$ for all t during each UL interval.*

3. Outline of the proof

We follow [10] quite closely. As before, we establish the limit recursively, considering each successive UL and OL interval in turn, with these intervals being determined by the fluid model. The values of $B_n(t, y)$ and $B(t, y)$ at the final time t of each interval serve as the initial values at the beginning of the next interval. We assume that we start in a UL interval. All intervals after the first necessarily begin critically loaded. In the neighborhood of each switching point, the number of customers in the system can oscillate above and below the staffing level. It is necessary to show that the impact of the critical loading at the switching points is asymptotically negligible, but that is already done in Sections 6.1 and 8 of [10]; the same reasoning applies here. In [10], it was shown that the limit in the FCLT is affected by the switching (the difference being expressed via the two processes $\hat{X}$ and $\hat{X}^*$), but not the FWLLN. The FWLLN is unaffected here as well.

The result for UL intervals is relatively elementary, because we can apply the two-parameter heavy-traffic limits for infinite-server models in [14] together with a new two-parameter limit for the remaining fluid that was in the system initially at the beginning of the interval; see Section 4.

As in the previous two papers [9,10], the analysis during each OL interval depends on a careful analysis of the flow of customers or fluid into service. For *GI* service, the fluid model involves a fixed point equation for, $b(t,0)$, the rate fluid enters service, (18) in Section 6.2 of [9], namely,

$$b(t,0) = a(t) + \int_0^t b(t-x,0)g(x)\,dx,$$
$$a(t) \equiv \dot{s}(t) + \tau(b,g,t), \tag{4}$$

for $\tau(b,g,t)$ in (2), with $\dot{s}(t)$ being the derivative of s.

Just as in [10], for each OL interval, we use the compactness approach to establish our FWLLN; see Section 11.6 of [20]. We prove that the sequence of FWLLN-scaled stochastic processes $\{\bar{B}_n(t,0)\}$ is C-tight in $\mathbb{D}$; see Theorem 11.6.3 of [20]. That implies that every subsequence has a further converging subsequence with a continuous limit. We then prove convergence by showing that all convergent subsequences have the same limit. The critical step is to show that, asymptotically, the flow into service satisfies the fixed point equation (4), which is shown to have a unique solution in Theorem 2 of [9] by virtue of the Banach contraction fixed point theorem. That theorem implies that all limits of convergent subsequences must be the same, and thus coincide with the fluid model in [9].

Since the system is overloaded, the flow into service is determined by the creation of newly available capacity through departures and changes in the staffing level. It is easy to see that the sequence of scaled departure processes $\{\bar{D}_n\}$ is C-tight in $\mathbb{D}$. From the C-tightness of the sequence of departure processes we easily obtain C-tightness for the associated sequence $\{\bar{B}_n(t,0)\}$. We next apply the limit for the sequence of processes specifying the fluid content in service at time 0 that remains in service later from Section 4. We then apply [14] to obtain a limit along the convergent subsequence of the scaled process recording the new customers to enter service. With all those results, we can show that the limits of these convergent subsequences must satisfy the fixed point equation (4).

Once a limit has been obtained for the process describing the flow of customers into service, we obtain associated limits for the queue-length and waiting-time processes exactly as in our previous paper [10]; nothing new is required for those limits.

4. A fluid limit for the customers initially in service

We consider a single UL or OL interval, which we take to begin at time 0, with the initial number of customers that have been in service for a duration at most y being $B_n(0,y)$, $y \geq 0$. Let $B_n^o(t,y)$ be the number of customers in service at time t that have been in service for a duration of at most y, from among those in service at time 0. Hence, $B_n^o(0,y) = B_n(0,y)$. Let the customers initially in service be ordered according to their ages in service. Let $\tau_{n,i}$ denote the length of time customer i has been in service at time 0 in model n, so that $0 \leq \tau_{n,1} \leq \tau_{n,2} \leq \cdots$. With that convention, we can write

$$B_n^o(t,y) = \sum_{i=1}^{B_n(0,y-t)} 1\{\eta_i(\tau_{n,i}) \geq t\}, \tag{5}$$

where, conditional on the sequence $\{\tau_{n,i}: i \geq 1\}$, the sequence $\{\eta_i(\tau_{n,i}): i \geq 1\}$ is a sequence of independent random variables, with

$$P(\eta_i(x) > t) \equiv H_x^c(t) \equiv \frac{G^c(t+x)}{G^c(x)}, \quad x \geq 0, t \geq 0, \tag{6}$$

where G is the service-time cdf with $G^c(x) \equiv 1 - G(x) > 0$ for all x.

Theorem 2 (*FWLLN for Old Service Content*)**.** *During the OL or UL interval starting at time 0, if*

$$\bar{B}_n(0,y) \Rightarrow B(0,y) = \int_0^y b(0,x)\,dx \quad \text{in } \mathbb{D} \text{ as } n \to \infty, \tag{7}$$

and H_x^c is defined in (6), then

$$\bar{B}_n^o(t,y) \Rightarrow B^o(t,y) \equiv \int_0^{(y-t)^+} b(0,x)H_x^c(t)\,dx \quad \text{in } \mathbb{D}_{\mathbb{D}} \text{ as } n \to \infty. \tag{8}$$

Proof. Add and subtract $H_{\tau_{n,i}}^c(t) \equiv G^c(t+\tau_{n,i})/G^c(\tau_{n,i})$ inside the sum (5):

$$\begin{aligned}\bar{B}_n^o(t,y) &= n^{-1}\sum_{i=1}^{B_n(0,y-t)} \left(1\{\eta_i(\tau_{n,i}) \geq t\} - H_{\tau_{n,i}}^c(t)\right) \\ &\quad + n^{-1}\sum_{i=1}^{B_n(0,y-t)} H_{\tau_{n,i}}^c(t) \\ &= n^{-1}\sum_{i=1}^{B_n(0,y-t)} \left(1\{\eta_i(\tau_{n,i}) \geq t\} - H_{\tau_{n,i}}^c(t)\right) \\ &\quad + \int_0^{y-t} H_x^c(t)\,d\bar{B}_n(0,x).\end{aligned} \tag{9}$$

The first term in (9). We consider the two terms on the right in (9) in turn, starting with the first, denoted by $\bar{B}_{n,1}^o$. Conditional on any possible realization of the sequence $\{\tau_{n,i}: i \geq 1\}$, for each fixed t and y, $\bar{B}_{n,1}^o(t,y)$ is a scaled random sum of independent mean-zero two-point random variables, each taking values within the interval $[-1,1]$. Together with condition (7), that implies the convergence $\bar{B}_{n,1}^o(t,y) \Rightarrow 0$ for each fixed t and y, by virtue of the law of large numbers for non-identically distributed triangular arrays, e.g., Theorem 1 on p. 316 of [3]. We show that convergence is uniform in the following lemma, proved in Section 7. Let $\|\cdot\|$ be the uniform norm.

Lemma 1. *The convergence of the first term in (9) is uniform in t and y; i.e., $\|\bar{B}_{n,1}^o(\cdot,\cdot)\| \Rightarrow 0$.*

The second term. Now consider the second term on the right in (9), denoted by $\bar{B}_{n,2}^o$. Observe that $H_x^c(t) \equiv G^c(t+x)/G^c(x)$ is uniformly continuous in t and x over the relevant finite intervals $0 \leq t \leq T$ and $0 \leq x \leq y^{\uparrow}$, because we have assumed that G^c is continuous and $G^c(x) > 0$ for all x. Hence, for any $\epsilon > 0$, there exists k and $0 \equiv t_0 < t_1 < \cdots < t_k \equiv T$, such that $|H_x^c(t) - H_x^c(t_k)| < \epsilon$ for $t_{k-1} \leq t \leq t_k$ and all x, $0 \leq x \leq T + y^{\uparrow}$. Hence it suffices to focus on only finitely many t_i.

Notice that $\bar{B}_n(0,y)$ is nondecreasing in y, and so can be regarded as a random cdf associated with a random measure. Thus the integral $\bar{B}_{n,2}^o(t,y)$ for each fixed t and y is a continuous map, as in (2.1) on p. 77 of [20]. We thus can apply the continuous mapping theorem to establish the convergence $\bar{B}_{n,2}^o(t,y) \Rightarrow B(t,y)$ for each fixed t and y. However, since the integrand is nonnegative, $\bar{B}_{n,2}^o(t,\cdot)$ is also a finite random measure. Since the limit $B(t,\cdot)$ is continuous, the convergence of cdf's is uniform in y. Moreover, continuity extends directly to the k time points t_i above. Hence, we can apply the continuous mapping theorem to get $\{\bar{B}_{n,2}^o(t_i,\cdot): 1 \leq i \leq k\} \Rightarrow \{B^o(t_i,\cdot): 1 \leq i \leq k\}$ in $\mathbb{D}^k$ as $n \to \infty$, for $B^o(t_i,y)$ in (8). Hence, with the approximation in the first paragraph, we have deduced that $\|\bar{B}_{n,2}^o - B^o\| \Rightarrow 0$, so that we can ignore the second term, and that completes the proof. □

5. The UL intervals

As indicated in Section 3, we treat the successive UL and OL intervals recursively, starting with an initial UL interval. The switching times are defined by the limiting fluid model. The limit

for the scaled content in service at the final time of each interval provides the limit for the initial scaled content in service at the beginning of the next interval. As in [10], without loss of generality, it suffices to focus on each interval with the initial time set at 0.

With Theorem 2, the limiting behavior during each UL interval is elementary. We can treat the new input separately from the old content. The new content during a UL interval can be treated as in the associated $G_t/GI/\infty$ infinite-server model, applying the two-parameter limit in [14]. The new content is then independent of the old content. The fluid limit is then the sum of the two. We thus obtain a FWLLN for the sequence of processes $\{\bar{B}_n: n \geq 1\} \equiv \{\{n^{-1}B_n(t,y): t \geq 0, y \geq 0\}: n \geq 1\}$ in the function space $\mathbb{D}_{\mathbb{D}}$. The limit is the fluid model during a UL interval, exactly as described in [9]. Since, there are no customers waiting during a UL interval, there are no other processes to consider.

A complication is that the limiting fluid model starts critically loaded at time 0 for all intervals considered except perhaps the initial interval, and ends critically loaded at the interval endpoint when the system shifts to OL. As explained in Section 3 of [10], for each UL interval $[\tau_1, \tau_2]$, we require that the total fluid content in the system, $X(t)$, satisfy $X(t) < s(t)$ for all t such that $\tau_1 < t < \tau_2$. That implies, for each $\epsilon > 0$ with $2\epsilon < \tau_2 - \tau_1$, that the queueing system must eventually be underloaded in the interval $[\tau_1+\epsilon, \tau_2-\epsilon]$. However, the queueing systems can be temporarily overloaded in the neighborhoods of the endpoints. Nevertheless, that does not alter the FWLLN. The supporting argument given in Sections 6.1 and 8 of [10] applies here as well.

6. The OL intervals

The OL intervals are more complicated for two reasons. First, there is a queue, so that there is waiting and abandonment. Second, even to only describe the content in service, the fluid model requires solving the fixed point equation (4) to determine the rate fluid enters service. As indicated in Section 3, the queue, waiting and abandonment can be treated just as in [10], once we have established the convergence of the LLN-scaled processes that count the number of customers entering service. (The only difference is that $b(t,0) = \dot{s}(t) + \mu s(t)$ for M service, whereas $b(t,0)$ solves the fixed point Eq. (4) for GI service.) Thus, for the FWLLN with GI service, we only need to establish a limit for the scaled two-parameter process describing the number of customers in service with specified durations. We outlined the proof in Section 3; we now provide the details.

The entering-service and departure processes. The sequence of scaled departure processes $\{\bar{D}_n\}$ is C-tight in $\mathbb{D}$, using the characterization in Theorem 11.6.3 of [20], because it has a time-dependent and state-dependent rate at any time that is bounded above by a constant, the product of the suprema of the staffing level and the service hazard rates over finite intervals, allowing stochastic bounds using a constant rate Poisson process. The associated LLN-scaled number of customers to have entered service in the interval $[0,t]$, $\bar{E}_n(t) \equiv \bar{B}_n(t,t)$, satisfies

$$\begin{aligned}\bar{E}_n(t) &= n^{-1}(s_n(t) - s_n(0)) + \bar{D}_n(t) + o(1) \quad \text{as } n \to \infty,\\ &= s(t) - s(0) + \bar{D}_n(t) + o(1) \quad \text{as } n \to \infty,\end{aligned}$$

where the $o(1)$ term on the first line accounts for the asymptotically negligible contribution contributed by the consequence of the system being only critically loaded at time 0; see Section 6.1 of [10]. As an immediate consequence of the C-tightness of $\{\bar{D}_n\}$ and the smoothness assumption about the deterministic staffing function, we see that the sequence of processes $\{\bar{E}_n\}$ is C-tight in $\mathbb{D}$ as well.

Now we express $\bar{E}_n(t)$ in terms of the scaled departure process of old customers, $\bar{D}_n^o(t)$, and the scaled departure process of new customers to arrive in the interval $[0,t]$, $\bar{D}_n^v(t)$. In particular, we write

$$\bar{E}_n(t) = n^{-1}(s_n(t) - s_n(0)) + \bar{D}_n^o(t) + \bar{D}_n^v(t) + o(1) \quad \text{as } n \to \infty. \tag{10}$$

First, the departures of old customers can be represented as $\bar{D}_n^o(t) = \bar{B}_n^o(0) - \bar{B}_n^o(t)$. Thus, from Section 4, we can deduce that $\bar{D}_n^o(t) \Rightarrow B^o(0) - B^o(t)$ in $\mathbb{D}$.

Second, we see that the departure process of new customers has the same mathematical form as the departure process from an infinite-server queue with arrival process $E_n(t)$ and service times distributed as G; i.e.,

$$\bar{D}_n^v(t) \equiv n^{-1}\sum_{i=1}^{E_n(t)} \mathbf{1}(A_i^{(n)} + S_i \leq t),$$

where $A_1^{(n)}, A_2^{(n)}, \ldots$ are arrival times associated with counting process E_n, $S_1, S_2, \ldots$ are i.i.d. service times, each having cdf G.

Since the sequence of processes $\{\bar{E}_n\}$ is C-tight in $\mathbb{D}$, every subsequence has a convergent subsequence. So consider some convergent subsequence. Without introducing special notation for subsequences, suppose that $\bar{E}_n \Rightarrow E$ in $\mathbb{D}$. By Theorem 3.1 of [14], $\bar{D}_n^v(t) \Rightarrow D^v(t)$ in $\mathbb{D}$, where

$$D^v(t) = \int_0^t G(t-s)dE(s), \quad t \geq 0. \tag{11}$$

From the C-tightness of the sequence $\{\bar{E}_n\}$, we deduce that each limit function of the convergent subsequence, E, must be Lipschitz continuous as well as nondecreasing, which implies that E is differentiable almost everywhere with respect to Lebesgue measure. Hence, the limit E of the converging subsequence can be represented as $E(t) = \int_0^t e(s)\,ds$ for $e(s) = b(s,0)$.

Combining the results above, we obtain convergence to an integral equation, i.e.,

$$\bar{E}_n(t) \Rightarrow E(t) = B^o(0) - B^o(t) + \int_0^t G(t-s)e(s)ds \quad \text{as } n \to \infty, \tag{12}$$

where the convergence holds by the continuous mapping theorem using the function of summation. As a consequence, by differentiating (12), we see that the derivative of the limit of every convergent subsequence of $\{\bar{E}_n\}$ must satisfy the fixed point equation (4). By Theorem 2 of [9], this equation has a unique solution. Hence all convergent subsequences must have the same limit. Hence we must have full convergence to the fluid function in [9].

The two-parameter service content process. We next establish the convergence of the two-parameter function $\bar{B}_n(t,y)$. We divide into two terms: (i) old customers $\bar{B}_n^o(t,y)$ and (ii) new customers $\bar{B}_n^v(t,y)$. Since the convergence of $\bar{B}_n^o(t,y)$ is already obtained in Theorem 2, it only remains to treat $\bar{B}_n^v(t,y)$. By (12) and Theorem 3.1 of [14],

$$\begin{aligned}\bar{B}_n^v(t,y) &= \frac{1}{n}\sum_{i=E_n(t-y)+1}^{E_n(t)} \mathbf{1}(A_i^{(n)} + S_i > t)\\ &\Rightarrow \int_{t-y}^t G^c(t-s)dE(s) \equiv B^v(t,y)\end{aligned}$$

in $\mathbb{D}_{\mathbb{D}}$. Therefore,

$$\begin{aligned}\bar{B}_n(t,y) &= \bar{B}_n^o(t,y) + \bar{B}_n^v(t,y) \Rightarrow B^o(t,y) + B^v(t,y) \equiv B(t,y)\\ &= \int_0^{(y-t)^+} b(0,x)H_x^c(t)\,dx\\ &\quad + \int_{t-y}^t G^c(t-s)dE(s)\cdot\mathbf{1}(t>y).\end{aligned}$$

The two-parameter queue content process. As stated before, the FWLLN limit for the scaled queue-length process $\bar{Q}_n(t,y)$ follows by the argument given in [10]. However, that proof was stated

only for the one-parameter process $\bar{Q}_n(t) \equiv \bar{Q}_n(t,\infty)$. Since the reasoning is the same for the two-parameter FWLLN, and straightforward given the two-parameter FWLLN in [14], we omit the lengthy details.

7. Proof of Lemma 1

As a first step, we will apply the martingale FCLT in Section 7.1 of [2] to establish the convergence

$$\hat{X}_n^o(t,x) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nx \rfloor} \left(1\{\eta_i(\tau_{n,i}) \geq t\} - H^c_{\tau_{n,i}}(t) \right) \Rightarrow \mathcal{B}(\sigma_t^2(x)) \quad (13)$$

in $\mathbb{D}([0,y^\uparrow],\mathbb{R})$ for each fixed t, where $\mathcal{B}$ is a standard Brownian motion and $\{\sigma_t^2(x): x \geq 0\}$ is a deterministic variance function, to be specified below.

To justify (13), it is convenient to work with the point process $Z_n(k) \equiv \tau_{n,k}$ instead of the counting process $B_n(0,y) \equiv \min\{k \geq 0: \tau_{n,k} \leq y\}$. The scaled versions of these processes are related asymptotically by the inverse map in Sections 13.6–13.8 of [20]. The inverse process associated with the scaled counting process $\{\bar{B}_n(0,y): y \geq 0\}$ is $\{\bar{Z}_n(y): y \geq 0\}$, where $\bar{Z}_n(y) \equiv n^{-1}Z_n(\lfloor ny \rfloor)$; see Lemma 13.6.6 of [20]. To put this properly in the setting of [20], we can modify all these processes and functions to make them unbounded above. That can be done by setting $B(0,x^\uparrow + x) \equiv B(0,x^\uparrow) + x$ for $x > 0$ and similarly for the processes $B_n(0,y)$. Then convergence of the original processes will follow by restricting to a finite interval.

Applying the continuous mapping theorem with the inverse function with (7), we obtain $\bar{Z}_n(y) \Rightarrow Z(y)$ in $\mathbb{D}$, where $Z(y)$ is the inverse of $B(0,y)$, i.e., $B(0,Z(y)) = y$ for all y. Specifically, we apply Theorem 13.6.2 of [20], recalling that $B(0,\cdot)$ is strictly increasing and continuous over the interval $[0,x^\uparrow]$ by Assumption 1, so that all the Skorohod topologies reduce to uniform convergence. Moreover, $Z(y) = \int_0^y z(x)\,dx$ where $z(x) = 1/b(0,Z(x))$ by the inverse function theorem from calculus.

To determine the variance function $\sigma_t^2(x)$ in (13), we observe that, conditional on any possible value for the sequence $\{\tau_{n,i}: i \geq 1\}$,

$$\begin{aligned} \mathrm{Var}(\hat{X}_n^o(t,y)) &= \frac{1}{n}\sum_{i=1}^{\lfloor nx \rfloor} \mathrm{Var}(\eta_i(\tau_{n,i})) = \frac{1}{n}\sum_{i=1}^{\lfloor nx \rfloor} H_{\tau_{n,i}}(t) H^c_{\tau_{n,i}}(t) \\ &= \int_0^x H_u(t) H_u^c(t)\, d\bar{Z}_n(u). \end{aligned} \quad (14)$$

Hence, unconditioning, we obtain

$$\begin{aligned} \mathrm{Var}(\hat{X}_n^o(t,x)) &= E\left[\int_0^x H_u(t) H_u^c(t)\, d\bar{Z}_n(u) \right] \\ &\Rightarrow \int_0^x H_u(t) H_u^c(t)\, dZ(u) \\ &= \int_0^x H_u(t) H_u^c(t) z(u)\, du \equiv \sigma_t^2(x), \end{aligned} \quad (15)$$

because the integrand is a bounded continuous function. In particular, we can apply the Skorohod representation theorem to replace convergence in distribution $\bar{Z}_n(y) \Rightarrow Z(y)$ in $\mathbb{D}$ with convergence w.p.1. Since $\bar{Z}_n$ is nonnegative and nondecreasing, the almost sure convergence $\bar{Z}_n(y) \to Z(y)$ in $\mathbb{D}$ corresponds to the almost sure convergence of finite measures over each bounded interval, which implies the limit in (15).

To justify the limit (13) for each fixed t, we first condition on the sequence $\{\tau_{n,i}: i \geq 1\}$ and then uncondition, as in Section 7.3 of [13]. Hence, we initially place the entire sequence $\{\tau_{n,i}\}$ in the filtration together with the natural filtration of the stochastic process $\{\hat{X}_n^o(t,x): x \geq 0\}$, with t fixed. The martingale FCLT, Theorem 1.4 (b) in Section 7.1 of [2], applies because, with that conditioning, the summands are independent bounded mean-zero random variables. Hence, the variance function $\{\mathrm{Var}(\hat{X}_n^o(t,x)): x \geq 0\}$ in (14) is the predictable quadratic variation process of the martingale with respect to the specified filtration, which converges as $n \to \infty$ to $\sigma_t^2(x)$ in (15). Moreover, the regularity conditions (1.16) and (1.17) on p. 340 of [2] hold.

However, we make no direct use of the FCLT or the variance in (15) here. Instead, we apply the FCLT for each fixed t to obtain the associated FWLLN after scaling further by dividing by $\sqrt{n}$. Since the limit in the FWLLN is the deterministic function $0e$, where $e(t) \equiv t$, $t \geq 0$, the FWLLN extends immediately to the joint distributions for each finite subset of t; see Theorem 11.4.5 of [20]; i.e.,

$$\left(\bar{X}_n^o(t_1,\cdot),\ldots,\bar{X}_n^o(t_k,\cdot)\right) \Rightarrow (0e,\ldots,0e) \quad \text{in } \mathbb{D}^k \text{ as } n \to \infty,$$

where $\bar{X}_n^o(t,x) \equiv n^{-1/2}\hat{X}_n^o(t,x)$ for $\hat{X}_n^o(t,x)$ in (13). It thus remains to establish tightness of the sequence $\{\bar{X}_n^o(t,\cdot): t \geq 0\}$ in $\mathbb{D}_\mathbb{D}$; see Section 6.2 of [14].

First observe that $\bar{X}_n^o(\cdot,x)$ is the scaled sum of $\lfloor nx \rfloor$ stochastic processes, each of which takes values in the interval $[-1,1]$. As a consequence, $|\bar{X}_n^o(t,x)| \leq x$ for all $n \geq 1$, $t \geq 0$ and $x \geq 0$. So the processes $\bar{X}_n^o(t,x)$ are uniformly bounded.

Given the uniform boundedness, we can apply the stopping-time characterization of tightness in $\mathbb{D}_\mathbb{D}$ in Remark 6.1 of [14]; see also Condition 1° on p. 176 of [1]. That is, we will show that, for any $\epsilon > 0$ and $\eta > 0$ and any bounded stopping time τ with respect to the filtration, that there exists $\delta > 0$ and n_0 such that

$$P\left(\|\bar{X}_n^o(\tau+\delta,\cdot) - \bar{X}_n^o(\tau,\cdot)\| > \eta \right) < \epsilon \quad (16)$$

for all $n \geq n_0$.

The desired property (16) follows from the structure of the summands. Observe that $\bar{X}_n^o(\cdot,x)$ is the scaled sum of $\lfloor nx \rfloor$ stochastic processes, each of which is a uniformly continuous deterministic function except for a single discontinuity of size 1, which occurs at a random time. Except for the discontinuities, for any $\epsilon > 0$, there exists $\delta > 0$ such that the oscillation over each interval $[t, t+\delta]$ is less than ϵ uniformly over t and all the component sample paths, provided that no discontinuity is encountered. For tightness, the discontinuities are the critical part.

Fortunately, we control the occurrence of these discontinuities in the component processes being summed, because the probabilities that discontinuities fall in any interval can be bounded. In particular, we have

$$P(t_1 \leq \eta_x \leq t_2) \leq g^\uparrow |t_2 - t_1| / G^c(T + y^\uparrow), \quad (17)$$

uniformly in all x under consideration, where $g^\uparrow \equiv \sup_{0 \leq x \leq T+y^\uparrow} g(x)$. Thus, for any stopping time, the number of service completions among these customers initially in the interval $[\tau, \tau+\delta]$ is bounded above by the sum of $y^\uparrow n$ i.i.d random variables with probabilities governed by the bound in (17), which is $c\delta$ for a constant c. The contribution from terms with no arrivals in this interval is easily seen to be of order $O(\delta)$. Hence, $\delta > 0$ can be chosen to achieve (16). That proves the tightness, and thus the convergence $\bar{X}_n^o(\cdot,\cdot) \Rightarrow X^o(\cdot,\cdot)$ in $\mathbb{D}_\mathbb{D}$, where $X^o(t,y) = 0$ for all t, y.

Given the FWLLN for the two-parameter process $\bar{X}_n^o(t,y)$, we treat the random sum itself by applying the continuous mapping theorem with the composition map, using (7) together with the FWLLN for $\bar{X}_n^o(\cdot,\cdot)$. In this two-parameter setting we can apply Theorem 2.4 of [18].

Acknowledgment

The second author received support from NSF grant CMMI 1066372.

References

[1] P. Billingsley, Convergence of Probability Measures, second ed., Wiley, New York, 1999.

[2] S.N. Ethier, T.G. Kurtz, Markov Processes: Characterization and Convergence, Wiley, New York, 1986.
[3] W. Feller, An Introduction to Probability Theory and its Applications, John Wiley and Sons, New York, 1971.
[4] L.V. Green, P.J. Kolesar, W. Whitt, Coping with time-varying demand when setting staffing requirements for a service system, Prod. Oper. Manag. 16 (2007) 13–39.
[5] W. Kang, K. Ramanan, Fluid limits of many-server queues with reneging, Ann. Appl. Prob. 20 (2010) 2204–2260.
[6] H. Kaspi, K. Ramanan, Law of large numbers limits for many-server queues, Ann. Appl. Prob. 21 (2011) 33–114.
[7] H. Kaspi, K. Ramanan, SPDE limits of many-server queues. arxiv:1010.0330v1, October 2, 2010.
[8] E.V. Krichagina, A.A. Puhalskii, A heavy-traffic analysis of a closed queueing system with a GI/∞ service center, Queueing Syst. 25 (1997) 235–280.
[9] Y. Liu, W. Whitt, The $G_t/GI/s_t + GI$ many-server fluid queue, Queueing Systems. http://dx.doi.org/10.1007/s11134-012-9291-0 (forthcoming).
[10] Y. Liu, W. Whitt, Many-server heavy-traffic limit for queues with time-varying parameters, Columbia University, NY, 2011. http://www.columbia.edu/~ww2040/allpapers.html (submitted for publication).
[11] Y. Liu, W. Whitt, A network of time-varying many-server fluid queues with customer abandonment, Oper. Res. 59 (2011) 835–846.
[12] A. Mandelbaum, W.A. Massey, M.I. Reiman, Strong approximations for Markovian service networks, Queueing Syst. 30 (1998) 149–201.
[13] G. Pang, R. Talreja, W. Whitt, Martingale proofs of many-server heavy-traffic limits for Markovian queues, Probab. Surv. 4 (2007) 193–267.
[14] G. Pang, W. Whitt, Two-parameter heavy-traffic limits for infinite-server queues, Queueing Syst. 65 (2010) 325–364.
[15] A.A. Puhalskii, J. Reed, On many-server queues in heavy traffic, Ann. Appl. Prob. 20 (2010) 129–195.
[16] J. Reed, The $G/GI/N$ queue in the Halfin–Whitt regime, Ann. Appl. Prob. 19 (2009) 2211–2269.
[17] A.V. Skorohod, Limit theorems for stochastic processes, Theor. Probab. Appl. 1 (1956) 261–290.
[18] R. Talreja, W. Whitt, Heavy-traffic limits for waiting times in many-server queues with abandonment, Ann. Appl. Prob. 19 (2009) 2137–2175.
[19] W. Whitt, Fluid models for multiserver queues with abandonments, Oper. Res. 54 (2006) 37–54.
[20] W. Whitt, Stochastic-Process Limits, Springer, New York, 2002.
[21] J. Zhang, Fluid models of many-server queues with abandonment, Queueing Syst., http://dx.doi.org/10.1007/s11134-012-9307-9 (forthcoming).