

# The $G/GI/N$ Queue in the Halfin-Whitt Regime II: Idle Time System Equations

J. E Reed

Department of Information, Operations, and Management Sciences  
Leonard N. Stern School of Business  
New York University

December 14, 2007

## Abstract

In this paper, we study the  $G/GI/N$  queue in the Halfin-Whitt regime from the point of view of the servers in the system. The main results obtained in this paper are identical to those found in [11], however, the manner in which they are obtained is entirely different. In particular, rather than modeling the  $G/GI/N$  queue similarly to the  $G/GI/\infty$  queue as in [11], we model the queue length process as the difference between the number of arrivals to and the number of departures from the system. This approach requires a centering of the departure process from the system which is rather unconventional in its nature and one of our first results is to show that this centering is in fact correct. We then proceed to provide both fluid and diffusion limits for the queue length process of the  $G/GI/N$  queue in the Halfin-Whitt regime.

AMS 2000 Subject Classification 60G15, 60G44, 60K25

Keyword and Phrases queueing theory, diffusion approximation, Gaussian process, martingale, weak convergence, central limit theorem for renewal processes

## 1 Introduction

This paper is a continuation of our study of the  $G/GI/N$  queue in the Halfin-Whitt regime which was initiated in [11]. In [11], a general methodology labeled the “Infinite Server Queue System Equations” approach was provided for obtaining fluid and diffusion limits for the queue length process of the  $G/GI/N$  queue in the Halfin-Whitt regime. This approach relied heavily upon keeping track of the amount of time each customer arriving to the system has to wait before receiving service in conjunction with some results based off the heavy-traffic theory for infinite server queues. In the end, our main results in [11] show that in the limit, as the number of servers grows large, the  $G/GI/N$  queue in the Halfin-Whitt regime behaves similarly to an infinite sever

queue in heavy traffic with a slight adjustment term necessary to keep track of the amount of time that customers are forced to wait.

In the present paper, we take the opposite approach to that in found in [11] and provide a second, entirely distinct yet equally compelling methodology for proving the main results found in [11]. In particular, in the present paper we consider the approach of keeping track of the amount of time that each server in the system is forced to wait between serving customers. We label this approach the “Idle Time System Equations” approach and note that it just as easily could have been provided first in the prequel with the “Infinite Server Queue System Equations” approach appearing next. In the future, depending upon the application at hand, one may find is more or less fruitful to apply the “Infinite Server System Equations” approach versus the “Idle Time System Equations” approach.

The main idea behind the “Idle Time System Equations” approach is to directly use a conservation of flow approach in order to express the number of customers in the system as the difference between the number of arrivals to and the number departures from the system. The attentive reader will note that such an approach is what is many times encountered in conventional heavy traffic analysis, see, for instance, [7], [12], and [5]. However, several difficulties arise in the application of such an approach to the setting of the unconventional Halfin-Whitt regime which are not typically encountered in the more traditional, conventional heavy-traffic regime. The most significant of these difficulties lies in the fact that in the conventional heavy-traffic regime one is able to rely on the functional central limit for renewal processes in conjunction with the random time change theorem in order to show convergence of the centered departure process from the system to a limiting Brownian motion. Unfortunately, in the setting of the unconventional Halfin-Whitt regime, such a direct approach is no longer possible. Thus, a secondary contribution of the present paper is to provide an efficient way around this problem and interested reader is referred to Section 3 to see how these difficulties are resolved.

In order to provide some further background and also some additional motivation for the results found in this paper, we begin by providing a summary of the results found in [11]. In Section 5 of [11], the author considers a sequence of  $G/GI/N$  queues indexed by the number of servers  $N$  where the service time distribution is held fixed across  $N$  with CDF  $F$ . The arrival rate to the  $N^{th}$  system increases with  $N$  in such a way that the overall traffic intensity of the system converges to one at a rate which is proportional to the square root of the number of servers. Specifically, defining the traffic intensity of the  $N^{th}$  system to be the quantity  $\rho^N$ , the author makes the the Halfin-Whitt (or Q.E.D.) assumption

$$\sqrt{N}(1 - \rho^N) \rightarrow \beta \text{ as } N \rightarrow \infty,$$

where  $-\infty < \beta < \infty$ .

The main result of Section 5 of [11] is to provide a heavy-traffic limit for the diffusion scaled queue length process for the  $G/GI/N$  queue in the above described Halfin-Whitt regime. Let  $Q^N = \{Q^N(t), t \geq 0\}$  be the queue length process of the  $N^{th}$  system described in the previous

paragraph (here, “queue length” is being used to refer to the total number of customers in the system), and letting

$$\tilde{Q}^N(t) = \frac{Q^N(t) - N}{\sqrt{N}}, \quad t \geq 0,$$

set  $\tilde{Q}^N = \{\tilde{Q}^N(t), t \geq 0\}$  to be the diffusion scaled queue length process in the  $N^{\text{th}}$  system. Then, under appropriate initial conditions on the system at time 0-, Theorem 2 of [11] states that  $\tilde{Q}^N$  converges weakly as  $N$  tends to  $\infty$  to a limiting stochastic process,  $\tilde{Q}_F$ , which may be best summarized as the unique solution to a stochastic convolution equation (see uppermost arrow in Figure 1 below). Specifically, one has the following.

**Theorem 2 of [11].** *If the initial residual service time distribution  $F_0 = F_e$  and  $\tilde{Q}_0^N \Rightarrow \tilde{Q}_0$  as  $N \rightarrow \infty$ , then  $\tilde{Q}^N \Rightarrow \tilde{Q}_F$  as  $N \rightarrow \infty$ , where  $\tilde{Q}_F$  is the unique strong solution to the stochastic convolution equation*

$$\tilde{Q}_F(t) = \tilde{M}_Q(t) + \tilde{Q}_I(t) - \beta F_e(t) + \int_0^t \tilde{Q}_F^+(t-s) dF(s), \quad \text{for } t \geq 0, \quad (1.1)$$

where  $\tilde{Q}_F^+ = \max(\tilde{Q}_F, 0)$ ,  $F$  is CDF of the service time distribution,  $F_e$  is the equilibrium distribution associated with  $F$  (see (6.4)) and  $\tilde{M}_Q$  is an additional process which is related to the initial conditions of the queue.

Note that one may interpret Theorem 1 as a precise statement of the fact that in limit as the number of servers tends to  $\infty$  in the Halfin-Whitt regime, the  $G/GI/N$  queue behaves similarly to an  $G/GI/\infty$  queue in a similarly defined heavy-traffic regime, thus providing further motivation for the label “Infinte Server Queue System Equations” approach. The interested reader is referred to the excellent papers [2, 6, 9] for a further of analysis of the  $G/GI/\infty$  queue in its heavy-traffic regime. It is also worthwhile to point out the presence of the term  $\tilde{Q}_F^+$  in the limit in (6.27). This term may be heuristically explained by the fact that our  $G/GI/N$  approximation to the  $G/GI/\infty$  queue will be off whenever there are some customers in the system who were forced to wait some amount of before entering service.

Motivated by the desire to show that in the case of exponentially distributed service times the results of [4] and [11] agree, a second characterization,  $\tilde{Q}_M$ , of the limiting process in Theorem 2 of [11] was also provided by Corollary 2 of [11]. This characterization was arrived at in [11] by a direct (sample-path) manipulation of the limiting process of Theorem 2 of [11] (see righthand arrow of Figure 1 above). We have the following from [11].

**Corollary 1 of [11].** *The limiting process,  $\tilde{Q}_F$ , of Theorem 1 is equivalent in distribution to  $\tilde{Q}_M$ , the unique strong solution to*

$$\tilde{Q}_M(t) = \tilde{\zeta}(t) + \int_0^t \tilde{\zeta}(t-s) dM(s) - \beta t - \int_0^t \tilde{Q}_M^-(t-s) dM(s), \quad t \geq 0, \quad (1.2)$$

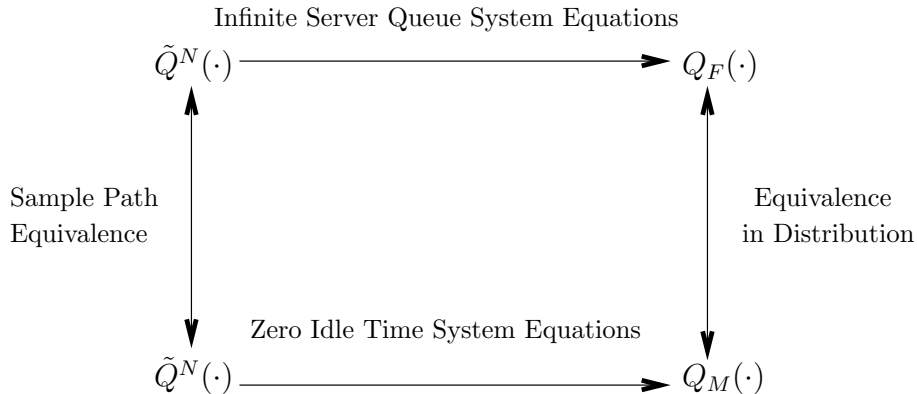


Figure 1: Commutation of the system equations for the  $G/GI/N$  queue.

where  $\tilde{\zeta} = \tilde{M}_Q + \tilde{Q}_I$ ,  $\tilde{Q}_M^- = \min(0, \tilde{Q}_M)$  and  $M$  is the renewal function associated with the pure renewal process with interarrival distribution  $F$ .

As stated above, the characterization provided by Corollary 2 of [11] of the limit of Theorem 2 of [11] may be used to show that in the case of exponential service times, the results of [4] and [11] agree. Specifically, in this case we have that  $M$  is the renewal function associated with a rate one Poisson process so that  $dM(t) = dt$ . The convolution in (1.2) then reduces to an ordinary integration with respect to Lebesgue measure and a minimal number of direct calculations may be applied to show that  $\tilde{Q}_M$  is in fact the diffusion process obtained by Halfin and Whitt [4]. While clearly appealing, this result also raises the interesting question of how in general the limit in (1.2) should correctly be interpreted.

In Theorem 2 of Section 2 of the present paper, the “Idle Time System Equations” approach is shown to provide a direct proof of Corollary 2 of [11] (see the lowermost arrow in Figure 1 below). This result may then be used as a basis for providing a convenient interpretation of the limit in (1.2). We also shown in Corollary 2 that Theorem 2 of [11] may be derived as a Corollary to Theorem 2 in the present paper (see, again, righthand arrow in Figure 1 above). This fact may then be used to provide justification behind the statement that the “Idle Time System Equation Approach” and the “Infinite Server Queue System Equation” approach stand on equal footing. The key step in our analysis in the present paper which differs from the more conventional heavy-traffic approach is in determining a proper centering for the departure process from the system. In the case of exponential service times, one may see how our choice centering reduces to that which has been previously employed.

The remainder of this paper is now organized as follows. In the following Section, we present the “Idle Time System Equations” approach for the  $G/GI/N$  queue. The primary object of interest in this Section is the rather unconventional choice of centering we choose for the departure process (see (2.5) and (2.3)). Next, in Section 3, we show that this choice of centering is in fact correct

in that our centered departure process has expected value zero for all  $t \geq 0$ . Section 4 provides a regulator map result upon which our weak convergence results hinge. In Section 5, the first of our two main results are presented. Specifically, we provide a deterministic fluid limit for the fluid scaled queue length process of the  $G/GI/N$  queue under general initial conditions. In future work, we expect that this fluid limit will prove fruitful in providing a transient analysis of the  $G/GI/N$  queue. In Section 6, after placing two assumptions on the initial conditions of the system at time 0– and providing several preliminary results, we state our second main result, Theorem 2, which provides a weak limit for the diffusion scaled queue length process. The weak limit we obtain in Theorem 2 is, as discussed above, identical to the limit given by Corollary 2 of [11]. We next proceed in Section 6 to provide a limit characterizing the amount of time that each server must wait between serving customers. Finally, we conclude Section 6 with Corollary 2, which shows that the limit obtained by Theorem 2 may be used to obtain the limit given by Theorem 2 of [11]. The present paper is concluded in Section 7 with directions for future research. The proofs of several of the results in Sections 4 and 6 may be also found in the Appendix.

## 1.1 Notation

In this subsection, we provide the notation which will be used for the remainder of the paper. Note that our notation is the same as that used in [11]. All random variables and stochastic processes are defined on a common probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  and stochastic processes are assumed to be measurable maps from  $(\Omega, \mathcal{F})$  to  $(D[0, \infty), \mathcal{D})$ , the space of right continuous functions with left limits. We denote by  $\mathcal{D}$  the Borel  $\sigma$ -algebra generated by the Skorohod  $J_1$  topology. We also denote by  $d_{J_1}$  the Skorohod metric on  $D[0, \infty)$  and for each  $x \in D[0, \infty)$  and  $T \geq 0$ , set

$$\|x\|_T = \sup_{0 \leq t \leq T} |x(t)|$$

to be the uniform metric, which we sometimes denote by  $u$ . Finally, we denote the Euclidian metric on  $\mathbb{R}$  by  $|\cdot|$ . The notation  $\Rightarrow$  is used to denote converge in distribution.

For any two metric spaces  $(\mathcal{S}_1, m_1)$  and  $(\mathcal{S}_2, m_2)$ , we denote by  $(\mathcal{S}_1 \times \mathcal{S}_2, m_1 \times m_2)$  the product metric space where, for  $(x_1, x_2), (y_1, y_2) \in \mathcal{S}_1 \times \mathcal{S}_2$ , we set  $(m_1 \times m_2)((x_1, x_2), (y_1, y_2)) = \max\{m_1(x_1, y_1), m_2(x_2, y_2)\}$  to be the maximum metric. For each  $k \geq 1$ , we denote the product metric space  $(D[0, \infty) \times \dots \times D[0, \infty), d_{J_1} \times \dots \times d_{J_1})$  by  $(D^k[0, \infty), d_{J_1}^k)$  and the product metric space  $(D[0, \infty) \times \dots \times D[0, \infty), u \times \dots \times u)$  by  $(D^k[0, \infty), u^k)$ .

## 2 Idle Time System Equations

In this Section, we provide the “Idle Time System Equations” for the  $G/GI/N$  queue. As stated in the Introduction, the key difference between the approach detailed here and that given in the prequel [11] is that in this paper the system equations for the  $G/GI/N$  queue are written using a

conservation of flow approach as the difference between the number of arrivals and the number of departures whereas in [11], the  $G/GI/N$  queue was modeled similar to an infinite server queue.

Initially, at time  $0-$ , we assume that there are  $Q_0$  customers in the system. Since there are  $N$  servers in the system, the first  $\min(Q_0, N)$  of these  $Q_0$  initial customers will already have completed some service by time zero. We therefore denote the residual service times of those customers already in service by time zero by the i.i.d. sequence of random variables  $\{\tilde{\eta}_i, i \geq 1\}$  with common distribution  $F_0$ . Strictly for notational convenience in future Sections, we also assume that the first  $\min(Q_0, N)$  initial customers in service are being served by servers 1 through  $\min(Q_0, N)$ .

Customer arrive to the system according to the general arrival process  $A = \{A(t), t \geq 0\}$  and in general it is allowed that  $A(0) > 0$  and also possibly that two arrivals occur at the same time, i.e.  $A(t) > A(t-) + 1$ . Each customers arriving to the system are served on a first come first served basis (we assume that those customers waiting in the queue at time zero had some already determined order of arrival) and the service times of those customers entering service after time  $0-$  are assumed to be independent and identically distributed. Specifically, the service time of the  $i^{th}$  customer to enter service after time  $0-$  is given by  $\eta_i$  where we assume that  $\{\eta_i, i \geq 1\}$  is an i.i.d. sequence of mean 1 random variables with common distribution  $F$ .

Now, for each  $i = 1, \dots, N$ , let  $S_i(t)$  be the number of customers processed by server  $i$  in its first  $t \geq 0$  units of processing time and denote by  $S_i = \{S_i(t), t \geq 0\}$  the departure process from server  $i$ . In the case that server  $i$  had a customer in service at time  $0-$  (which, in our notation, corresponds to  $1 \leq i \leq \min(Q_0, N)$ ), it is clear by the i.i.d. assumption on the service times above that  $S_i$  will be a delayed renewal process with initial delay distribution given by the residual service time distribution  $F_0$  and subsequent interarrival distribution given by the service time distribution  $F$ . On the other hand, if server  $i$  was empty at time  $0-$ , then  $S_i$  will simply be a pure renewal process with interarrival distribution given by the service time distribution  $F$ .

Now letting  $B_i(t)$  be equal to the cumulative busy time of server  $i$  up until time  $t$  and setting  $B_i = \{B_i(t), t \geq 0\}$  to be the busy time process for server  $i$ , we have that the total number of customers in the system at time  $t \geq 0$  is given by

$$Q(t) = Q_0 + A(t) - \sum_{i=1}^N S_i(B_i(t)). \quad (2.1)$$

Note that equation (2.1) is the conservation of flow approach which one sees as the starting point in many heavy traffic analysis, see for instance [7],[12] and [5]. The one difference perhaps is the fact that in (2.1) we are summing the departure processes over multiple servers as opposed to a single server. This difference will however become significant in the Sections that follow where the number of servers will be taken to infinity.

The next step following (2.1) in conventional heavy traffic analysis is to “center” the arrival and departure processes. For the arrival process,  $A$ , such a centering is standard and we choose to center the arrival process by its rate process  $N\rho e = \{N\rho t, t \geq 0\}$ . We therefore define  $\hat{A} = A - N\rho e$  to be the centered arrival process. However, as already noted to in the Introduction, our centering

for the departure process becomes a highly non-trivial task as it is actually comprised of a sum of individual departure processes, one from each of the servers. In what follows, we provide our choice of centering for each of the individual servers. Later, in Section 3, we provide the intuition behind this choice of centering as well as a derivation that is in fact correct.

We begin by providing our choice of centering those servers which were idle at time  $0-$ . Let  $M = \{M(t), t \geq 0\}$  be the renewal function associated with the pure renewal process  $N = \{N(t), t \geq 0\}$  with interarrival distribution given by the service time distribution  $F$ . By definition (see Section 3.2 of [13]),  $M(t)$  is the expected number of renewals of the renewal process  $N$  by time  $t$ . Also note that by Exercise 3.4 of [13],  $M$  is given by the unique solution to the renewal equation

$$M(t) = F(t) + \int_0^t F(t-s)dM(s), \quad t \geq 0. \quad (2.2)$$

Next, let  $Q_i(t)$  be equal to 1 if a customer is in service at server  $i$  at time  $t$  and let  $Q_i(t) = 0$  if not. We now define the centered departure process for servers  $i = \min(Q_0, N) + 1$  through  $N$  by

$$\hat{S}_i(t) = S_i(B_i(t)) - \left( M(t) - \int_0^t 1\{Q_i(t-s) = 0\}dM(s) \right), \quad t \geq 0, \quad (2.3)$$

and we set  $\hat{S}_i = \{\hat{S}_i(t), t \geq 0\}$ . Note that since servers  $i = \min(Q_0, N) + 1, \dots, N$  were idle at time  $0-$  we have that the process  $S_i$  is a pure renewal process with interarrival distribution  $F$ . However, the process  $S_i \circ B_i = \{S_i(B_i(t)), t \geq 0\}$  is clearly not a pure renewal which leads to the rather unconventional centering in (2.3) above. The intuition as well as some further explanation behind the choice of centering in (2.3) is provided in the Section that follows. However, for the moment we simply point out that in the case of mean 1 exponential service times,  $M(t) = t, t \geq 0$ , and hence, the reader may independent verify via the non-idling condition

$$B_i(t) = \int_0^t 1\{Q_i(s) = 1\}ds, \quad t \geq 0,$$

that  $\hat{S}_i(t) = S_i(B_i(t)) - B_i(t)$ , which is what has appeared before in previous work before on the  $GI/M/N$  queue.

We now give our choice of centering for the departure processes for servers  $i = 1, \dots, \min(Q_0, N)$  which were busy serving a customer at time  $0-$ . Let  $M_D = \{M_D(t), t \geq 0\}$  be the renewal function for a delayed renewal process  $N_D = \{N_D(t), t \geq 0\}$  with initial delay distribution given by the initial residual service time distribution  $F_0$  and with interarrival distribution given by the service time distribution  $F$ . By definition (see for instance, Section 3.3. of [13]),  $M_D(t) = E[N_D(t)], t \geq 0$ . Furthermore, by Exercise 3.25, part a, of [13], it follows that  $M_D$  is the given by

$$M_D(t) = F_0(t) + \int_0^t F_0(t-s)dM(s), \quad t \geq 0. \quad (2.4)$$

Again, let  $Q_i(t) = 1$  if a customer is in service at server  $i$  at time  $t \geq 0$  and  $Q_i(t) = 0$  if not. We then define the centered departure processes for servers  $i = 1$  through  $\min(Q_0, N)$  by

$$\hat{S}_i(t) = S_i(B_i(t)) - \left( M_D(t) - \int_0^t 1\{Q_i(t-s) = 0\} dM(s) \right), \quad t \geq 0, \quad (2.5)$$

and we set  $\hat{S}_i = \{\hat{S}_i(t), t \geq 0\}$ . Note that servers  $i = 1, \dots, \min(Q_0, N)$  represent those servers which were busy serving a customer at time  $0-$ . Hence, for such servers, the process  $S_i$  is a delayed renewal process with initial delay distribution given by the residual service time distribution  $F_0$  and interarrival distribution given by the service time distribution  $F$ . However, as above, the process  $S_i \circ B_i = \{S_i(B_i(t)), t \geq 0\}$  will not in general turn out to be delayed renewal process which again leads to the rather unconventional centering in (2.5) above.

We now complete this Section by substituting the centered arrival process and departure processes into the basic equation (2.1) for the queue length process. First note that by substituting the centered departure processes (2.3) and (2.5) into equation (2.1), we obtain, after some algebra,

$$\begin{aligned} Q(t) &= Q_0 + A(t) - \hat{S}(t) - \min(Q_0, N)M_D(t) - (N - Q_0)^+ M(t) \\ &\quad + \sum_{i=1}^N \int_0^t 1\{Q_i(t-s) = 0\} dM(s), \end{aligned} \quad (2.6)$$

where we define the centered departure process from the system to be

$$\hat{S}(t) = \sum_{i=1}^N \hat{S}_i(t), \quad t \geq 0, \quad (2.7)$$

and we set  $\hat{S} = \{\hat{S}(t), t \geq 0\}$ . The process  $\hat{S}$  is the sum of the centered departure processes from each of the individual servers.

We next treat the summation appearing in (2.6) which arises from the centering of the departure processes. First recall that since we are operating under a non-idling discipline, it follows that for  $t \geq 0$ ,

$$\text{number of idle servers at time } t = \sum_{i=1}^N 1\{Q_i(t) = 0\} = -(Q(t) - N)^-,$$

where  $(Q(t) - N)^- = \min(0, (Q(t) - N))$ .

Thus,

$$\begin{aligned} \sum_{i=1}^N \int_0^t 1\{Q_i(t-s) = 0\} dM(s) &= \int_0^t \sum_{i=1}^N 1\{Q_i(t-s) = 0\} dM(s) \\ &= - \int_0^t (Q(t-s) - N)^- dM(s), \end{aligned} \quad (2.8)$$

and so, substituting (2.8) into equation (2.6) for the queue length process, we obtain

$$\begin{aligned}
Q(t) &= Q_0 + A(t) - \hat{S}(t) - \min(Q_0, N)M_D(t) - (N - Q_0)^+M(t) \\
&\quad - \int_0^t (Q(t-s) - N)^- dM(s).
\end{aligned} \tag{2.9}$$

Equation (2.9) will be useful for our analysis in the Sections that follow. Most importantly, it allows us to use the continuous mapping approach which is typical in many heavy-traffic proofs. However, we postpone providing a regulator map result upon which our continuous mapping approach is based until Section 4 and instead, choose to study in more detail the centered departure processes appearing in (2.3) and (2.5) in the Section that follows.

### 3 Centering the Departure Processes

In this Section, we study the centered departure processes appearing in (2.3) and (2.5) in more detail. In particular, we provide justification for the choices of centering appearing in (2.3) and (2.5). This justification was intentionally left out of Section 2 in order to allow a more efficient presentation of our choice of centering and also because we feel that the analysis presented here may be of independent interest. We begin by providing alternate representations for the centered departure processes in terms of a certain empirical distribution process. This representation is then helpful in showing that these processes are in fact properly centered as well as in deriving our fluid and diffusion limit results in Sections 5 and 6.

Let

$$S(t) = \sum_{i=1}^N S_i(B_i(t)), \quad t \geq 0,$$

be the total number of departures from the system by time  $t \geq 0$  and set

$$d_k = \inf\{t \geq 0 : S(t) \geq k\} \tag{3.1}$$

to be the time of the  $k^{\text{th}}$  departure from the system. Furthermore, denote by  $\tilde{v}_k$  the amount of time that the  $k^{\text{th}}$  idle server at time 0– must waiting before receiving its first customer and by  $v_k$  the amount of time that the server which is responsible for  $k^{\text{th}}$  departure from the system must wait before receiving its next customer. Finally, let  $\eta_k$  denote the service time of the  $k^{\text{th}}$  customer to enter service after time 0–. We then have that  $S(t)$ , the total number of departures from the system by time  $t \geq 0$ , may be expressed as

$$S(t) = \sum_{k=1}^{\min(Q_0, N)} 1\{\tilde{\eta}_k \leq t\} + \sum_{k=1}^{(N-Q_0)^+} 1\{\tilde{v}_k + \eta_k \leq t\} + \sum_{k=1}^{S(t)} 1\{d_k + v_k + \eta_k \leq t\}, \tag{3.2}$$

for  $t \geq 0$ .

We will now choose to center each of the processes appearing in (3.2) above by their means conditional on their departure times and server idle times. By doing so, we obtain

$$N(t) = \sum_{k=1}^{\min(Q_0, N)} (1\{\tilde{\eta}_k \leq t\} - F_0(t)) \quad (3.3)$$

and

$$\begin{aligned} S_F(t) &= \sum_{k=1}^{(N-Q_0)^+} (1\{\tilde{v}_k + \eta_k \leq t\} - F(t - \tilde{v}_k)) \\ &\quad + \sum_{k=1}^{S(t)} (1\{d_k + v_k + \eta_k \leq t\} - F(t - v_k - \eta_k)), \end{aligned} \quad (3.4)$$

for  $t \geq 0$ .

We now provide a result which will be crucial in the proof of our main result of this Section which shows that centered departure is in fact properly centered. We have the following which may be viewed as an analogue to Proposition 1 of [11] for the “Idle Time System Equations”.

**Proposition 1** *For each  $t \geq 0$ ,*

$$\begin{aligned} &\sum_{k=1}^{S(t)} (F(t - d_k - v_k) - F(t - d_k)) \\ &= \int_0^t (Q(t - s) - N)^- dF(s) - \sum_{k=1}^{(N-Q_0)^+} (F(t - \tilde{v}_k) - F(t)). \end{aligned}$$

**Proof.** First note that at each time  $t \geq 0$ , the total number of idle servers is given by

$$(N - Q(t))^+ = \sum_{k=1}^{(N-Q_0)^+} 1\{\tilde{v}_k < t\} - \sum_{k=1}^{S(t)} 1\{d_k \leq t < d_k + v_k\}.$$

It therefore follows that

$$\begin{aligned} &\sum_{k=1}^{S(t)} (F(t - d_k - v_k) - F(t - d_k)) \\ &= \sum_{k=1}^{S(t)} \int_{(t-d_k)^+}^{t-v_k-d_k} dF(s) \\ &= - \sum_{k=1}^{S(t)} \int_0^\infty 1\{t - (d_k + v_k) < s \leq t - d_k\} dF(s) \end{aligned}$$

$$\begin{aligned}
&= - \sum_{k=1}^{S(t)} \int_0^\infty 1\{d_k \leq t-s < d_k + v_k\} dF(s) \\
&= - \int_0^\infty \sum_{k=1}^{S(t)} 1\{d_k \leq t-s < d_k + w_k\} dF(s) \\
&= - \int_0^t \left( (N - Q(t-s))^+ - \sum_{k=1}^{(N-Q_0)^+} 1\{\tilde{v}_k < t-s\} \right) dF(s) \\
&= \int_0^t (Q(t-s) - N)^- dF(s) + \int_0^t \sum_{k=1}^{(N-Q_0)^+} 1\{\tilde{v}_k < t-s\} dF(s).
\end{aligned}$$

A reverse argument can now also be used to show that

$$\int_0^t \sum_{k=1}^{(N-Q_0)^+} 1\{\tilde{v}_k < t-s\} dF(s) = - \sum_{k=1}^{(N-Q_0)^+} (F(t - \tilde{v}_k) - F(t)),$$

which completes the proof.  $\square$

We may now use Proposition 1 in order to provide a representation of the centered departure process  $\hat{S} = \{\hat{S}(t), t \geq 0\}$  in terms of the processes  $N = \{N(t), t \geq 0\}$  and  $S_F = \{S_F(t), t \geq 0\}$  defined above. Recall the definition of  $M = \{M(t), t \geq 0\}$  as the renewal function of the pure renewal process with interarrival distribution given by the service time distribution  $F$ . We then have the following.

**Proposition 2** For each  $t \geq 0$ ,

$$\hat{S}(t) = N(t) + S_F(t) + \int_0^t N(t-s) dM(s) + \int_0^t S_F(t-s) dM(s).$$

**Proof.** First note that by the definitions (2.3) and (2.5) for the centered departure processes and the relationship (2.8) for the idle time process, we have that for each  $t \geq 0$  the centered departure process may be expressed as

$$\hat{S}(t) = S(t) - \left( \min(Q_0, N) M_D(t) + (N - Q_0)^+ M(t) - \int_0^t (Q(t-s) - N)^- dM(s) \right). \quad (3.5)$$

Next, by (3.2), (3.3) and (3.4) we have, after a little bit of algebra, that, for each  $t \geq 0$ ,

$$S(t) = N(t) + S_F(t) + \min(Q_0, N) F_0(t) + \sum_{k=1}^{(N-Q_0)^+} F(t - \tilde{v}_k) + \sum_{k=1}^{S(t)} F(t - v_k - d_k). \quad (3.6)$$

Furthermore, by Proposition 1 it follows that

$$\begin{aligned} \sum_{k=1}^{S(t)} F(t - d_k - v_k) &= \int_0^t S(t-s) dF(s) + \int_0^t (Q(t-s) - N)^- dF(s) \\ &\quad - \sum_{k=1}^{(N-Q_0)^+} (F(t - \tilde{v}_k) - F(t)), \end{aligned} \quad (3.7)$$

for  $t \geq 0$ , where the first expression on the righthand side above follows since, integrating by parts, we have

$$\sum_{k=1}^{S(t)} F(t - d_k - v_k) = \int_0^t F(t-s) dS(s) = \int_0^t S(t-s) dF(s).$$

Substituting (3.7) into (3.6) we now obtain that

$$\begin{aligned} S(t) &= N(t) + S_F(t) + \min(Q_0, N)F_0(t) + (N - Q_0)^+ F(t) + \int_0^t S(t-s)F(s) \\ &\quad + \int_0^t (Q(t-s) - N)^- dF(s). \end{aligned} \quad (3.8)$$

Next, note that by (2.2) and (2.4) we have the relationships

$$F(t) - M(t) = - \int_0^t M(t-s) dF(s)$$

and

$$\begin{aligned} F_0(t) - M_D(t) &= - \int_0^t F_0(t-s) dM(s) \\ &= - \int_0^t F_0(t-s) d \left( F(s) + \int_0^s F(s-u) dM(u) \right) \\ &= - \int_0^t F(t-s) d \left( F_0(s) + \int_0^s F_0(s-u) dM(u) \right) \\ &= - \int_0^t F(t-s) dM_D(s) \\ &= - \int_0^t M_D(t-s) dF(s), \end{aligned}$$

for  $t \geq 0$ . Thus, substituting (3.8) into the righthand side of equation (3.5) for the centered departure process, it follows after a little bit of algebra that

$$\begin{aligned}\hat{S}(t) &= N(t) + S_F(t) - \int_0^t (S(t-s) - \min(Q_0, N)M_D(t-s) + (N - Q_0)^+M(t-s) \\ &\quad + \int_0^{t-s} (Q((t-s) - u) - N)^-dM(u))dF(s),\end{aligned}$$

which, by (3.5), is equivalent to

$$\hat{S}(t) = N(t) + S_F(t) - \int_0^t \hat{S}(t-s)dF(s), \quad (3.9)$$

for  $t \geq 0$ .

However, (3.9) may be viewed as an renewal equation in terms of  $N + S_F = \{N(t) + S_F(t), t \geq 0\}$ . However, since  $N + S_F \in D[0, \infty)$  is  $\mathbb{P}$ -a.s. locally bounded, it follows that the solution to (3.9) is given by

$$\hat{S}(t) = N(t) + S_F(t) + \int_0^t (N(t-s) + S_F(t-s))dM(s),$$

for  $t \geq 0$ , which completes the proof. □

The representation given by Proposition 2 of the centered departure process in terms of the processes  $N = \{N(t), t \geq 0\}$  and  $S_F = \{S_F(t), t \geq 0\}$  will be useful in proving our main fluid and diffusion limit results. However, as we note next it is also useful in showing that the centered departure process does in fact have expected value equal to 0 for each  $t \geq 0$ .

To see this, first note that by conditioning on the initial number of customers in the system at time  $0-$ ,  $Q_0$ , and, since the initial residual service time sequence  $\{\eta_i, i \geq 1\}$  is i.i.d. with common distribution  $F_0$  and independent of  $Q_0$ , it is clear by the definition of  $N(t)$  in (3.3) that for each  $t \geq 0$ , we have  $E[N(t)] = 0$ . We next show that the same holds true for  $S_F(t)$  for each  $t \geq 0$ .

**Proposition 3** *For each  $t \geq 0$ ,  $E[S_F(t)] = 0$ .*

**Proof.** First note that

$$\begin{aligned}E \left[ \sum_{k=1}^{(N-Q_0)^+} (1\{\tilde{v}_k + \eta_k \leq t\} - F(t - \tilde{v}_k)) | Q_0 \right] &= \sum_{k=1}^{(N-Q_0)^+} E[1\{\tilde{v}_k + \eta_k \leq t\} - F(t - \tilde{v}_k)] \\ &= 0,\end{aligned} \quad (3.10)$$

where the second equality above follows since

$$E[1\{\tilde{v}_k + \eta_k \leq t\} - F(t - \tilde{v}_k)] = E[E[1\{\tilde{v}_k + \eta_k \leq t\} - F(t - \tilde{v}_k) | \tilde{v}_k]]$$

and

$$E[1\{\tilde{v}_k + \eta_k \leq t\} - F(t - \tilde{v}_k)|\tilde{v}_k] = E[1\{\tilde{v}_k + \eta_k \leq t\}|\tilde{v}_k] - F(t - \tilde{v}_k) = 0.$$

Thus, by (3.10) it follows that

$$\begin{aligned} & E \left[ \sum_{k=1}^{(N-Q_0)^+} (1\{\tilde{v}_k + \eta_k \leq t\} - F(t - \tilde{v}_k)) \right] \\ &= E \left[ E \left[ \sum_{k=1}^{(N-Q_0)^+} (1\{\tilde{v}_k + \eta_k \leq t\} - F(t - \tilde{v}_k)) | Q_0 \right] \right] \\ &= 0. \end{aligned} \tag{3.11}$$

Next, first note that the final term on the righthand side of (3.2) may be equivalently written as

$$\sum_{k=1}^{S(t)} 1\{d_k + v_k + \eta_k \leq t\} = \sum_{k=1}^{\infty} 1\{d_k + v_k + \eta_k \leq t\}, \tag{3.12}$$

for  $t \geq 0$ . Next, for each  $t \geq 0$  and  $K \geq 0$ , we have that

$$\sum_{k=1}^K (1\{d_k + v_k + \eta_k \leq t\} - F(t - d_k - v_k)) \leq 2S(t).$$

Thus,

$$E \left[ \sum_{k=1}^K (1\{d_k + v_k + \eta_k \leq t\} - F(t - d_k - v_k)) \right] \leq 2E[S(t)] \leq 2N \max(M(t), M_D(t)).$$

Furthermore, since  $\max(M(t), M_D(t)) < \infty$ , it follows by the Bounded Convergence Theorem [3] that

$$\begin{aligned} & E \left[ \lim_{K \rightarrow \infty} \sum_{k=1}^K (1\{d_k + v_k + \eta_k \leq t\} - F(t - d_k - v_k)) \right] \\ &= \lim_{K \rightarrow \infty} E \left[ \sum_{k=1}^K (1\{d_k + v_k + \eta_k \leq t\} - F(t - d_k - v_k)) \right]. \end{aligned} \tag{3.13}$$

However, for each  $k \geq 1$ , by the i.i.d. assumption on the service time sequence  $\{\eta_k, k \geq 1\}$  and, since service times are independent of the previous departure time as well as the amount of time

that the server had to wait, it follows that

$$\begin{aligned}
E[1\{d_k + v_k + \eta_k \leq t\} - F(t - d_k - v_k)] &= E[1\{\eta_k \leq t - d_k - v_k\} - F(t - d_k - v_k)] \\
&= E[E[1\{\eta_k \leq t - d_k - v_k\} - F(t - d_k - v_k) | d_k, v_k]] \\
&= E[E[1\{\eta_k \leq t - d_k - v_k\} | d_k, v_k] - F(t - d_k - v_k)] \\
&= E[1\{\eta_k \leq t - d_k - v_k\} | d_k, v_k] - F(t - d_k - v_k) \\
&= 0,
\end{aligned}$$

so that

$$\begin{aligned}
E \left[ \sum_{k=1}^K (1\{d_k + v_k + \eta_k \leq t\} - F(t - d_k - v_k)) \right] &= \sum_{k=1}^K E [(1\{d_k + v_k + \eta_k \leq t\} - F(t - d_k - v_k))] \\
&= 0,
\end{aligned}$$

for each  $K \geq 0$ . Thus, by (3.13), we conclude that

$$E \left[ \sum_{k=1}^{\infty} (1\{d_k + v_k + \eta_k \leq t\} - F(t - d_k - v_k)) \right] = 0,$$

which, by (3.4), (3.11) and (3.12) completes the proof.  $\square$

We may now proceed to use the Proposition 3 and the representation for  $\hat{S}$  given by Proposition 2 in order to show that  $\hat{S}$  is in fact properly centered. Specifically, we have the following.

**Proposition 4** *For each  $t \geq 0$ ,  $E[\hat{S}(t)] = 0$ .*

**Proof.** By Proposition 2 and 3 and Fubini's Theorem [3], it follows that for each  $t \geq 0$ ,

$$\begin{aligned}
E[\hat{S}(t)] &= E \left[ N(t) + S_F(t) + \int_0^t N(t-s) dM(s) + \int_0^t S_F(t-s) dM(s) \right] \\
&= E[N(t)] + E[S_F(t)] + \int_0^t E[N(t-s)] dM(s) + \int_0^t E[S_F(t-s)] dM(s) \\
&= 0,
\end{aligned}$$

which completes the proof.  $\square$

Note that similar arguments to those above but for each individual server may also be used to show that for each  $i = 1, \dots, N$ , we have that  $E[\hat{S}_i(t)]$  for each  $t \geq 0$ . Thus, not only is the centered departure process from the system properly centered but the centered departure processes for each individual server is also properly centered as well.

We now complete this Section by providing alternative representations for the processes  $N$  and  $S_F$  defined in (3.3) and (3.4) above. These representations will be useful in proving our heavy traffic fluid and diffusion limit results in Section 5 and 6. For each  $k \geq 1$ , let

$$\tau_k = \inf\{t \geq 0 : A(t) \geq k\}$$

be the time of the  $k^{\text{th}}$  arrival to the system. Next, let  $w_k$  be the amount of time that the  $k^{\text{th}}$  customer to arrive to the system after time  $0-$  has to wait before being served and, for  $k = 1, \dots, (Q_0 - N)^+$ , define  $\tilde{w}_k$  to be the amount of time that  $(N + k)^{\text{th}}$  customer in the system at time  $0-$  has to wait before entering service. We then set for each  $t \geq 0$ ,

$$W_0(t) = \sum_{k=1}^{\min(Q_0, N)} (1\{\tilde{\eta}_k > t\} - (1 - F_0(t)))$$

and

$$\begin{aligned} M_2(t) &= \sum_{k=1}^{(Q_0 - N)^+} (1\{\tilde{w}_k + \eta_k > t\} - G(t - \tilde{w}_k)) \\ &\quad + \sum_{k=1}^{A(t)} (1\{\tau_k + w_k + \eta_{(Q_0 - N)^+ + k} > t\} - G(t - \tau_k - w_k)). \end{aligned}$$

Note that the processes  $W_0(t)$  and  $M_2(t)$  defined above were originally defined in Section 2 of [11]. Moreover, the results obtained in [11] regarding these processes will be useful in Section 5 and 6 as well. However, we note that the results which we will use from [11] do not need to be relied upon in our approach. We simply adopt them here for as a matter of convenience and for the sake of brevity. Specifically, the results which will use do not in any way rely upon the ‘‘Infinite Server Queue System Equations’’ approach. Thus, we point out, the two approaches remain entirely separate. For the moment, however, we have the following two results, which complete this Section.

**Proposition 5** *For each  $t \geq 0$ ,  $N(t) = -W_0(t)$ .*

**Proof.** We have, for  $t \geq 0$ ,

$$\begin{aligned} N(t) &= \sum_{k=1}^{\min(Q_0, N)} (1\{\tilde{\eta}_k \leq t\} - F_0(t)) \\ &= \sum_{k=1}^{\min(Q_0, N)} ((1\{\tilde{\eta}_k \leq t\} - 1) - (F_0(t) - 1)) \\ &= \sum_{k=1}^{\min(Q_0, N)} (-1\{\tilde{\eta}_k > t\} + (1 - F_0(t))) \\ &= -W_0(t), \end{aligned}$$

which completes the proof. □

We next establish a similar equivalency result between  $M_2(t)$  and  $S_F(t)$ .

**Proposition 6** *For each  $t \geq 0$ ,  $S_F(t) = -M_2(t)$ .*

**Proof.** We have, for  $t \geq 0$ ,

$$\begin{aligned}
S_F(t) &= \sum_{k=1}^{(N-Q_0)^+} (1\{\tilde{v}_k + \eta_k \leq t\} - F(t - \tilde{v}_k)) \\
&\quad + \sum_{i=1}^{S(t)} (1\{d_k + v_k + \eta_{(N-Q_0)^++i} \leq t\} - F(t - d_k - v_k)) \\
&= \sum_{i=1}^{(N-Q_0)^+} ((1\{\tilde{v}_k + \eta_k \leq t\} - 1) + (1 - F(t - \tilde{x}_i))) \\
&\quad + \sum_{i=1}^{S(t)} ((1\{d_k + v_k + \eta_{(N-Q_0)^++i} \leq t\} - 1) + (1 - F(t - d_k - v_k))) \\
&= \sum_{i=1}^{(N-Q_0)^+} (-1\{\tilde{v}_k + \eta_k > t\} + G(t - \tilde{x}_i)) \\
&\quad + \sum_{i=1}^{S(t)} (-1\{d_k + v_k + \eta_{(N-Q_0)^++i} > t\} + G(t - d_k - v_k)) \\
&= - \sum_{i=1}^{(N-Q_0)^+} (1\{\tilde{v}_k + \eta_k > t\} - G(t - \tilde{x}_i)) \\
&\quad - \sum_{i=1}^{S(t)} (1\{d_k + v_k + \eta_{(N-Q_0)^++i} > t\} - G(t - d_k - v_k)) \\
&= - \sum_{i=1}^{(Q_0-N)^+} (1\{\tilde{w}_k + \eta_k \geq t\} - G(t - \tilde{w}_k)) \\
&\quad - \sum_{i=1}^{A(t)} (1\{\tau_k + w_k + \eta_{(Q_0-N)^++i} \geq t\} - G(t - \tau_k - w_k)) \\
&= -M_2(t),
\end{aligned}$$

which completes the proof. □

## 4 A Regulator Map Result

In this Section, we provide a family of regulator maps which are useful in the proofs our main results, Theorems 1 and 2. Let us first denote by  $R$  the renewal function associated with a pure renewal process with interarrival distribution  $B$  and let  $a \in \mathbb{R}$ . Our main goal in this Section is to characterize solutions  $z \in D[0, \infty)$  to equations of the form

$$z(t) = x(t) - \int_0^t (z(t-s) + a)^- dR(s), \quad t \geq 0, \quad (4.1)$$

where  $x \in D[0, \infty)$  and  $c^- = \min(0, c)$  for  $c \in \mathbb{R}$ . We therefore define the mapping  $\psi_R^a : D[0, \infty) \mapsto D[0, \infty)$  to be such that  $\psi_R^a(x)$  is a solution to (4.1) for each  $x \in D[0, \infty)$ .

Recall also by Proposition 2 of [11], the definition of  $\varphi_B^a(x)$  as the unique  $z \in D[0, \infty)$  satisfying the integral equation

$$z(t) = x(t) + \int_0^t (z(t-s) + a)^+ dB(s), \quad t \geq 0, \quad (4.2)$$

where  $c^+ = \max(0, c)$  for  $c \in \mathbb{R}$  and  $B$  is a distribution function on  $\mathbb{R}$ . The following proposition now shows that  $\psi_R^a$  is uniquely defined and provides some regularity results for  $\psi_R^a$  as well as a representation of  $\psi_R^a$  in terms of the regulator map  $\varphi_B^a$ . Its proof may be found in the Appendix.

**Proposition 7** *For each  $x \in D[0, \infty)$ , there exists a unique solution  $\psi_R^a(x)$  to (4.1). The function  $\psi_R^a : D[0, \infty) \mapsto D[0, \infty)$  is Lipschitz continuous in the topology of uniform convergence over bounded intervals and measurable with respect to the Skorohod  $J_1$  topology. Moreover, letting  $B$  be the distribution function associated with the renewal function  $R$  and defining  $\vartheta_B^a : D[0, \infty) \mapsto D[0, \infty)$  by*

$$\vartheta_B^a(x)(t) = x(t) - \int_0^t (x(t-s) + a)^+ dB(s) + aB(t), \quad t \geq 0, \quad (4.3)$$

*we have the representation  $\psi_R^a = \varphi_B^a \circ \vartheta_B^a$ .*

## 5 Fluid Limit Results

In this Section, we obtain limiting results for the fluid scaled queue length process. In order to obtain these results, we consider a sequence of  $G/GI/N$  queues indexed by the number of servers  $N$ . The familiar reader will note that our setup and notation here is the same as in Section 3 of [11]. In particular, our convention is to use a superscript  $N$  to denote all quantities and processes associated with the  $N^{th}$  system in our sequence.

Initially, at time  $0-$ , we assume that there are  $Q_0^N$  in the  $N^{th}$  system. For  $i = 1, \dots, \min(Q_0^N, N)$ , we denote by  $\tilde{\eta}_i$  the residual service time of the  $i^{th}$  customer in service at time  $0-$ , where we

assume that  $\{\tilde{\eta}_i, i \geq 1\}$  is an i.i.d. sequence of random variables with common distribution  $F_0$ . Note that the initial residual service time distribution is not changing with the number of servers  $N$ . Furthermore, only the first  $\min(Q_0^N, N)$  customers in the system at time  $0-$  will actually be in service. The remaining  $(Q_0^N - N)^+$  customers will have to wait before being served.

Customers arrive to the  $N^{\text{th}}$  system according to the general arrival process  $A^N = \{A^N(t), t \geq 0\}$ . As in Section 2, in general we allow that  $A^N(0) > 0$  and also possibly that more than one arrival may occur at the same time, i.e.  $A^N(t) > A^N(t-)$ . The  $i^{\text{th}}$  customer to enter service in the  $N^{\text{th}}$  system after time  $0-$  is assigned the service time  $\eta_k$ . We assume that  $\{\eta_k, k \geq 1\}$  is an i.i.d. sequence of mean 1, random variables with common distribution  $F$ . Note that we place no restrictions on  $F$  beyond a first moment. Furthermore, the service time distribution is not changing with the number of servers  $N$ .

We now proceed to obtain our fluid limit results. This is accomplished as follows. First, we provide a convenient representation of the fluid scaled queue length process in the  $N^{\text{th}}$  system described above in terms of the regulator map defined in Proposition 7. Next, we state several preparatory lemmas concerning various components of the so-called “free” process associated with the regulator map. These results are then used to prove our main result of the Section, Theorem 1, which provides a deterministic convolution equation as the limit for the fluid scaled queue length process.

We begin by letting  $Q^N = \{Q^N(t), t \geq 0\}$  be the queue length process in the  $N^{\text{th}}$  system. Recall that by equation (2.9), we have, that for each  $t \geq 0$ ,

$$\begin{aligned} Q^N(t) &= Q_0^N + A^N(t) - \hat{S}^N(t) - \min(Q_0^N, N)M_D(t) - (N - Q_0^N)^+M(t) \\ &\quad - \int_0^t (Q^N(t-s) - N)^- dM(s). \end{aligned} \tag{5.1}$$

Thus, if we now define the fluid scaled quantities

$$\bar{Q}_0^N = \frac{Q_0^N}{N}, \tag{5.2}$$

$$\bar{Q}^N(t) = \frac{Q^N(t)}{N},$$

$$\bar{A}^N(t) = \frac{\hat{A}^N(t)}{N}, \tag{5.3}$$

and

$$\bar{S}^N(t) = \frac{\hat{S}_N(t)}{N},$$

for  $t \geq 0$ , we then have, dividing equation (5.1) by  $N$ , that

$$\begin{aligned} \bar{Q}^N(t) &= \bar{Q}_0^N + \bar{A}^N(t) - \bar{S}^N(t) - \min(\bar{Q}_0^N, 1)M_D(t) - (1 - \bar{Q}_0^N)^+M(t) \\ &\quad - \int_0^t (\bar{Q}^N(t-s) - 1)^- dM(s). \end{aligned} \tag{5.4}$$

Furthermore, since by Proposition 7 the map  $\psi_M^a$  is uniquely defined with  $M = \{M(t), t \geq 0\}$  being the renewal function associated with  $F$  and  $a = -1$ , it then follows from (5.4), after defining the fluid scaled processes

$$\begin{aligned}\bar{Q}^N &= \{\bar{Q}^N(t), t \geq 0\}, \\ \bar{A}^N &= \{\bar{A}^N, t \geq 0\}\end{aligned}$$

and

$$\bar{S}^N = \{\bar{S}^N(t), t \geq 0\}, \quad (5.5)$$

that

$$\bar{Q}^N = \psi_M^a(\bar{Q}_0^N + \bar{A}^N - \bar{S}^N - \min(\bar{Q}_0^N, 1)M_D - (1 - \bar{Q}_0^N)^+ M), \quad (5.6)$$

with  $a = -1$ .

The representation (5.6) will be used to prove the main result of this Section, Theorem 1. However, before stating our main result, we begin with the following several preparatory propositions. Our first result states the fluid scaled, centered departure process,  $\bar{S}^N$ , converges weakly to the zero process as the number of server  $N$  grows to  $\infty$ . This of course provides further justification for the choice centering in (2.3) and (2.5). Note also the lack of additional assumptions required for this proposition to hold.

**Proposition 8**  $\bar{S}^N \Rightarrow 0$  as  $N \rightarrow \infty$ .

**Proof.** First note that by Proposition 3, we have the representation, for  $t \geq 0$ ,

$$\bar{S}^N(t) = \bar{N}^N(t) + \bar{S}_F^N(t) + \int_0^t \bar{N}^N(t-s) dM(s) + \int_0^t \bar{S}_F^N(t-s) dM(s), \quad (5.7)$$

where  $\bar{N}^N(t) = N^{-1}N^N(t)$  and  $\bar{S}_F^N(t) = N^{-1}S_F^N(t)$ .

Next, recalling from Proposition 5 the relationship  $N^N(t) = -W_0^N(t)$ , it follows that

$$\bar{N}^N(t) + \int_0^t \bar{N}^N(t-s) dM(s) = -\bar{W}_0^N(t) - \int_0^t \bar{W}_0^N(t-s) dM(s), \quad t \geq 0,$$

where  $\bar{W}_0^N(t) = N^{-1}W_0^N(t)$ . However, by Proposition 3 of [11],  $\bar{W}_0^N \Rightarrow 0$  as  $N \rightarrow \infty$ , and thus, for each  $T \geq 0$ ,

$$\begin{aligned}\sup_{0 \leq t \leq T} \left| \bar{N}^N(t) + \int_0^t \bar{N}^N(t-s) dM(s) \right| &= \sup_{0 \leq t \leq T} \left| -\bar{W}_0^N(t) - \int_0^t \bar{W}_0^N(t-s) dM(s) \right| \\ &\leq (1 + M(T)) \sup_{0 \leq t \leq T} |\bar{W}_0^N(t)| \\ &\Rightarrow 0 \text{ as } N \rightarrow \infty.\end{aligned}$$

A similar proof using Proposition 6 and Proposition 4 of [11] also shows that for each  $T \geq 0$ ,

$$\sup_{0 \leq t \leq T} \left| \bar{S}_F^N(t) + \int_0^t \bar{S}_F^N(t-s) dM(s) \right| \Rightarrow 0 \text{ as } N \rightarrow \infty,$$

which, by (5.7), completes the proof.  $\square$

We are now ready to state the main result of this Section. This result provides a limit for the fluid scaled queue length process  $\bar{Q}^N$  as the number of servers tends to  $\infty$ . We have the following.

**Theorem 1** *If  $(\bar{Q}_0^N, \bar{A}^N) \Rightarrow (\bar{Q}_0, \bar{A})$  in  $(\mathbb{R} \times D[0, \infty), |\cdot| \times d_{J_1})$  as  $N \rightarrow \infty$ , where  $\bar{A}$  is a stochastic process with  $\mathbb{P}$ -a.s. continuous sample paths, then  $\bar{Q}^N \Rightarrow \bar{Q}$  as  $N \rightarrow \infty$ , where  $\bar{Q}$  is the unique strong solution to the convolution equation*

$$\bar{Q}(t) = \bar{Q}_0 + \bar{A}(t) - \min(\bar{Q}_0, 1)M_D(t) - (1 - \bar{Q}_0)^+ M(t) - \int_0^t \bar{Q}^-(t-s) dM(s), \quad (5.8)$$

for  $t \geq 0$ .

**Proof.** Letting

$$\check{A}^N = \bar{A}^N - \bar{S}^N,$$

it follows that

$$(\bar{Q}_0^N, \check{A}^N) = (\bar{Q}_0^N, \bar{A}^N) + (0, -\bar{S}^N). \quad (5.9)$$

By Proposition 8 and Theorem 3.9 of [1],  $(0, -\bar{S}^N) \Rightarrow (0, 0)$  as  $N \rightarrow \infty$ , and hence, by the assumption  $(\bar{Q}_0^N, \bar{A}^N) \Rightarrow (\bar{Q}_0, \bar{A})$  in  $(\mathbb{R} \times D[0, \infty), |\cdot| \times d_{J_1})$  as  $N \rightarrow \infty$ , where  $\bar{A}$  has  $\mathbb{P}$ -a.s. continuous sample paths, the representation (5.9) and the continuous mapping Theorem [1], it follows that

$$(\bar{Q}_0^N, \check{A}^N) \Rightarrow (\bar{Q}_0, \bar{A}) \text{ in } (\mathbb{R} \times D[0, \infty), |\cdot| \times d_{J_1}) \text{ as } N \rightarrow \infty.$$

Next, since by Theorem 11.4.1 in [14], the space  $\mathbb{R} \times D[0, \infty)$  is separable under the product topology induced by the maximum metric  $|\cdot| \times d_{J_1}$ , it follows by the Skorohod Representation Theorem [14] that there exists some alternate probability space,  $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$ , on which are defined a sequence of processes

$$\{(\hat{Q}_0^N, \hat{A}^N), N \geq 1\} \quad (5.10)$$

and a single process  $(\hat{Q}_0, \hat{A})$  such that

$$(\hat{Q}_0^N, \hat{A}^N) \stackrel{d}{=} (\bar{Q}_0^N, \check{A}^N), \quad N \geq 1 \quad (5.11)$$

and

$$(\hat{Q}_0, \hat{A}) \stackrel{d}{=} (\bar{Q}_0, \bar{A}), \quad (5.12)$$

where

$$(\hat{Q}_0^N, \hat{A}^N) \rightarrow (\hat{Q}_0, \hat{A}) \text{ in } (\mathbb{R} \times D[0, \infty), |\cdot| \times d_{J_1}) \text{ } \hat{\mathbb{P}}\text{-a.s. as } N \rightarrow \infty. \quad (5.13)$$

Moreover, since the limit process  $\hat{A}$  is assumed to be continuous, it follows that the convergence in (5.13) may be strengthened to convergence in  $(\mathbb{R} \times D[0, \infty), |\cdot| \times u)$ .

Now, for each  $N \geq 1$  set

$$\hat{B}^N(t) = \min(\hat{Q}_0^N, 1)M_D(t) + (1 - \hat{Q}_0^N)^+M(t), \quad t \geq 0,$$

and

$$\bar{B}^N(t) = \min(\bar{Q}_0^N, 1)M_D(t) + (1 - \bar{Q}_0^N)^+M(t), \quad t \geq 0.$$

note that by (5.11) we have, setting  $\hat{B}^N = \{\hat{B}^N, t \geq 0\}$  and  $\bar{B}^N = \{\bar{B}^N, t \geq 0\}$ , that

$$\hat{B}^N \stackrel{d}{=} \bar{B}^N, \quad N \geq 1. \quad (5.14)$$

Furthermore, letting

$$\hat{B}(t) = \min(\hat{Q}_0, 1)M_D(t) + (1 - \hat{Q}_0)^+M(t), \quad t \geq 0, \quad (5.15)$$

it follows by (5.13) that for each  $T \geq 0$ ,

$$\begin{aligned} & \sup_{0 \leq t \leq T} |\hat{B}^N(t) - \hat{B}(t)| \quad (5.16) \\ &= \sup_{0 \leq t \leq T} |(\min(\hat{Q}_0^N, 1)M_D(t) + (1 - \hat{Q}_0^N)^+M(t)) - (\min(\hat{Q}_0, 1)M_D(t) + (1 - \hat{Q}_0)^+M(t))| \\ &= \sup_{0 \leq t \leq T} |(\min(\hat{Q}_0^N, 1) - \min(\hat{Q}_0, 1))M_D(t) + ((1 - \hat{Q}_0^N)^+ - (1 - \hat{Q}_0)^+)M(t)| \\ &\leq \sup_{0 \leq t \leq T} |(\min(\hat{Q}_0^N, 1) - \min(\hat{Q}_0, 1))M_D(t)| + \sup_{0 \leq t \leq T} |((1 - \hat{Q}_0^N)^+ - (1 - \hat{Q}_0)^+)M(t)| \\ &\leq M_D(T)|\min(\hat{Q}_0^N, 1) - \min(\hat{Q}_0, 1)| + M(T)|(1 - \hat{Q}_0^N)^+ - (1 - \hat{Q}_0)^+| \\ &\leq \max(M_D(T), M(T))(|\min(\hat{Q}_0^N, 1) - \min(\hat{Q}_0, 1)| + |(1 - \hat{Q}_0^N)^+ - (1 - \hat{Q}_0)^+|) \\ &\rightarrow 0 \text{ as } N \rightarrow \infty, \end{aligned}$$

and thus, letting  $\hat{B} = \{\hat{B}(t), t \geq 0\}$ , we have that  $\hat{B}^N \rightarrow \hat{B}$  in  $(D[0, \infty), u)$ ,  $\hat{\mathbb{P}}$ -a.s. as  $N \rightarrow \infty$ .

Now note that by (5.6), (5.11) and (5.14), setting

$$\hat{Q}^N = \psi_M^a(\hat{Q}_0^N + \hat{A}^N - \hat{B}^N), \quad N \geq 1,$$

with  $a = -1$ , we have by the measurability of  $\psi_M^a$  from Proposition 7 that

$$\hat{Q}^N \stackrel{d}{=} \bar{Q}^N, \quad N \geq 1. \quad (5.17)$$

Furthermore, since by (5.13) and (5.16),

$$\hat{Q}_0^N + \hat{A}^N - \hat{B}^N \rightarrow \hat{Q}_0 + \hat{A} - \hat{B} \text{ in } (D[0, \infty), u) \text{ } \hat{\mathbb{P}}\text{-a.s. as } N \rightarrow \infty,$$

it follows by the continuity portion of Proposition 7 that

$$\hat{Q}^N = \psi_M^a(\hat{Q}_0^N + \hat{A}^N - \hat{B}^N) \rightarrow \psi_M^a(\hat{Q}_0 + \hat{A} - \hat{B}) \text{ in } (D[0, \infty), u) \text{ } \hat{\mathbb{P}}\text{-a.s. as } N \rightarrow \infty.$$

Thus, since by (5.12) and the definition of  $\hat{B}$  in (5.15),

$$\hat{Q}_0 + \hat{A} - \hat{B} \stackrel{d}{=} \bar{Q}_0 + \bar{A} - \min(\bar{Q}_0, 1)M_D(t) + (1 - \bar{Q}_0)^+M(t),$$

we have by the measurability portion of Proposition 7 that  $\psi_M^a(\hat{Q}_0 + \hat{A} - \hat{B}) \stackrel{d}{=} \psi_M^a(\bar{Q}_0 + \bar{A} - \min(\bar{Q}_0, 1)M_D - (1 - \bar{Q}_0)^+M)$ , which, by the definition of  $\psi_M^a$  and (5.8) now completes the proof.  $\square$

## 6 Diffusion Limit Results

In this Section, we obtain limiting results for the diffusion scaled queue length process of the  $G/GI/N$  queue in the Halfin-Whitt regime. The results obtained in this Section are more sensitive to the stochasticity of the discrete system than the fluid limit results obtained in Section 5. We begin in Subsection 6.1 by defining the Halfin-Whitt heavy-traffic regime. Next, in Subsection 6.2, we specify the initial conditions of our system at time  $0-$  and, in Corollary 2, provide a limit for the fluid scaled queue length process under these initial conditions. In Subsection 6.3, we obtain our main result of the Section, Theorem 2, which provides a limit for the diffusion scaled queue length process as the number of servers goes to  $\infty$ . This limit may best be summarized as the unique solution to a stochastic convolution equation. Using Theorem 2, we then proceed in Proposition 10 of Subsection 6.4 to obtain a limit for the diffusion scaled virtual idle time process, which keeps track of the amount time that each server must wait before receiving its next customer. Finally, we conclude in Subsection 6.5 with Corollary 2, which provides an alternate representation of the limit obtained in Theorem 2 for the diffusion scaled queue length process. This alternate representation may be seen to be equivalent to the limit obtained in Theorem 2 of [11] using the ‘‘Infinite Server Queue System Equations’’ approach.

## 6.1 The Halfin-Whitt Heavy Traffic Regime

In this Subsection, we provide the details of the Halfin-Whitt heavy traffic regime. This regime may be considered a special instance of the fluid limit regime described in Section 5, with a few additional assumptions. The familiar reader will note that the regime described in this Subsection is identical to that described in Subsection 6.2 of [11].

Our underlying premise is that we are considering a sequence of  $G/GI/N$  queues which are indexed by the number of servers  $N$ . Initially, at time  $0-$ , we assume that there are  $Q_0^N$  customers in the  $N^{\text{th}}$  system. The first  $\min(Q_0^N, N)$  of these customers will already be in the process of being served by time  $0-$  and so we denote by  $\tilde{\eta}_K$  the residual service time of the  $K^{\text{th}}$  customer in service at time  $0-$ . We assume that  $\{\tilde{\eta}_K, K \geq 1\}$  is an i.i.d. sequence of random variables with common distribution  $F_0$ . Later in this Section, an exact specification of  $F_0$  will be given.

Customers arrive to the  $N^{\text{th}}$  system according to the arrival process  $A^N = \{A^N(t), t \geq 0\}$  and in general, as in Section 5, we allow the possibility that  $A^N(0) > 0$  and also that two or more arrivals may occur at the same time, i.e.  $A^N(t) > A^N(t-) + 1$ . Furthermore, we also assume that there exists a sequence of constants  $\{N\rho^N, N \geq 1\}$ , which may be loosely interpreted as the arrivals rates to the system, such that, setting

$$\tilde{A}^N(t) = \frac{A^N(t) - N\rho^N t}{\sqrt{N}}, \quad t \geq 0, \quad (6.1)$$

and  $\tilde{A}^N = \{\tilde{A}^N(t), t \geq 0\}$ , we have that

$$\tilde{A}^N \Rightarrow \tilde{\xi} \quad \text{as } N \rightarrow \infty, \quad (6.2)$$

where  $\tilde{\xi}$  is stochastic process with  $\mathbb{P}$ -a.s. continuous sample paths. The limiting process  $\tilde{\xi}$  given above captures the stochastic fluctuations of the arrival processes around their means. The interested reader is also referred to Section 2 of [11] for a more complete discussion of assumption (6.2) above.

The  $k^{\text{th}}$  customer to enter service in the  $N^{\text{th}}$  system after time  $0-$  is assigned the service time  $\eta_k, k \geq 1$ , where we assume that  $\{\eta_k, k \geq 1\}$  is an i.i.d. sequence of mean 1 random variables with common distribution  $F$ . We denote by  $G = 1 - F$  the tail distribution of  $F$ . Note that our sequence of customer service times is not changing with  $N$ . Furthermore, we place no additional assumptions on the service time distribution,  $F$ , beyond requiring a first moment.

In order to now complete our specification of the Halfin-Whitt regime, we must place an assumption on the behavior of the traffic intensity of the  $N^{\text{th}}$  system as  $N$  approaches  $\infty$ . First note that, since there are  $N$  servers in the  $N^{\text{th}}$  system, each with a service rate of 1, we have that the total capacity of the  $N^{\text{th}}$  system is given by  $N$ . Furthermore, if one loosely interprets  $N\rho^N$  as the arrival rate to the  $N^{\text{th}}$  system, it then follows that the traffic intensity of the  $N^{\text{th}}$  system is given by  $\rho^N$ . The Halfin-Whitt regime is now achieved by assuming that

$$\sqrt{N}(1 - \rho^N) \rightarrow \beta \quad \text{as } N \rightarrow \infty, \quad (6.3)$$

where  $-\infty < \beta < \infty$ . In words, the traffic intensity of the  $N^{\text{th}}$  system converges to one at a rate which is proportional to the number of servers in the system. This assumption will be crucial in the proof of our main result.

## 6.2 Initial Conditions

For the remainder of this Section, we make two simplifying assumptions on the initial conditions of the system at time  $0-$ . These assumptions are helpful to ensure that the limiting fluid scaled number of customers in the system remains constant for all time.

Our first assumption is that the fluid scaled number of customers in the system at time  $0-$  converges to one. That is,

$$\bar{Q}_0^N \Rightarrow 1 \text{ as } N \rightarrow \infty,$$

where we recall the definition of  $\bar{Q}_0^N$  from (5.2).

Our next assumption concerns the distribution of the residual service times of those customers in service at time  $0-$ . We assume that those customers in service at time zero have a residual service time distribution equal to  $F_e$ , the equilibrium distribution associated with  $F$ , which is defined by

$$F_e(x) = \int_0^x G(u)du, \quad x \geq 0. \quad (6.4)$$

Note that under this assumption, the delayed renewal process corresponding to initial delay distribution  $F_e$  and interarrival distribution  $F$  is stationary and that by Theorem 3.5.2 of [13] we have that the delayed renewal function  $M_D(t) = t$  for  $t \geq 0$ . Thus, this assumption is helpful in ensuring that the limiting fluid scaled departure rate from the system remains constant.

Under the very two special initial conditions given above, the limiting fluid scaled queue length process of Theorem 1 may be explicitly solved for and its solution is given by the zero process. Indeed, this result has already been proven in Corollary 1 of [11], however, we reprove it here again in order to demonstrate the usefulness of Theorem 1 of the current paper. The following is now the main result of this Subsection.

**Corollary 1** *If  $F_0 = F_e$  and  $\bar{Q}_0^N \Rightarrow 1$  as  $N \rightarrow \infty$ , then  $\bar{Q}^N \Rightarrow 1$  as  $N \rightarrow \infty$ .*

**Proof.** By the definition of  $\bar{A}^N$  in (5.3) and (5.5), assumption (6.2) and the Halfin-Whitt condition (6.3), we have the weak law of large numbers result,  $\bar{A}^N \Rightarrow e$  as  $N \rightarrow \infty$ . Furthermore, by the assumption  $\bar{Q}_1^N \Rightarrow 1$  and Theorem 3.9 of [1], we have the joint convergence  $(\bar{Q}_0^N, \bar{A}^N) \Rightarrow (1, e)$  in  $(\mathbb{R} \times D, \mathcal{B}(\mathbb{R}) \times \mathcal{D})$  as  $N \rightarrow \infty$ .

It now follows by Theorem 1 of Section 5 that  $\bar{Q}^N \Rightarrow \bar{Q}$  as  $N \rightarrow \infty$ , where  $\bar{Q}$  is given by the

unique strong solution to

$$\begin{aligned}
\bar{Q}(t) &= 1 + e(t) - \min(1, 1)M_D(t) - (1 - 1)^+M(t) - \int_0^t (\bar{Q}(t - s) - 1)^- dM(s) \\
&= 1 + t - M_D(t) - \int_0^t (\bar{Q}(t - s) - 1)^- dM(s) \\
&= 1 + t - t - \int_0^t (\bar{Q}(t - s) - 1)^- dM(s) \\
&= 1 - \int_0^t (\bar{Q}(t - s) - 1)^- dM(s),
\end{aligned}$$

for  $t \geq 0$ , which has the unique solution  $\bar{Q}(t) = 1$  for  $t \geq 0$ . This completes the proof.  $\square$

### 6.3 Weak Convergence Results

In this Subsection, we obtain our weak convergence result for the diffusion scaled queue length process, included in which is our second main result, Theorem 2. Our plan is to proceed in a manner similar to as in Section 5. In particular, we first provide a convenient representation for the queue length process in terms of the regulator map  $\psi_M^a$  given by Proposition 7. We then provide a crucial proposition on the joint convergence of the diffusion scaled initial number of customers in the system at time  $0-$  and the diffusion scaled, centered arrival and departure processes.. This proposition is useful in the proof of our main result. Finally, we conclude with our main result, Theorem 2, which provides the solution to a stochastic convolution equation as a limiting approximation for the diffusion scaled queue length process.

We begin by letting  $Q^N = \{Q^N(t), t \geq 0\}$  be the queue length process in the  $N^{\text{th}}$  system and setting

$$\tilde{Q}^N(t) = \frac{Q^N(t) - N}{\sqrt{N}}, \quad t \geq 0. \quad (6.5)$$

We also define  $\tilde{Q}^N = \{\tilde{Q}^N(t), t \geq 0\}$  be the diffusion scaled queue length process.

Now note that as pointed out in Subsection 6.2 above, under the assumption  $F_0 = F_e$  we have that the delayed renewal function  $M_D(t) = t, t \geq 0$ , and so algebraic manipulations of the queue length process in equation (2.9) yield

$$\begin{aligned}
(Q^N(t) - N) &= (Q_0^N - N) + \hat{A}^N(t) - \hat{S}^N(t) - N(1 - \rho^N)t + (Q_0^N - N)^-(M(t) - t) \\
&\quad - \int_0^t (Q^N(t - s) - N)^- dM(s).
\end{aligned} \quad (6.6)$$

Next, setting

$$\tilde{S}^N(t) = \frac{\hat{S}(t)}{\sqrt{N}}, \quad t \geq 0, \quad (6.7)$$

and defining  $\tilde{S}^N = \{\tilde{S}^N(t), t \geq 0\}$  to be the diffusion scaled service time process, it then follows, dividing equation (6.6) by  $\sqrt{N}$ , that

$$\begin{aligned} \tilde{Q}^N(t) &= \tilde{Q}_0^N + \tilde{Q}_0^{N,-} (M(t) - t) + \tilde{A}^N(t) - \tilde{S}^N(t) - \sqrt{N}(1 - \rho^N)e \\ &\quad - \int_0^t \tilde{Q}^{N,-}(t-s) dM(s), \end{aligned} \quad (6.8)$$

where we set

$$\tilde{Q}_0^N = \frac{Q_0^N - N}{\sqrt{N}}.$$

Thus, by (6.8) and the uniqueness portion of Proposition 7 of Section 4, we have that

$$\tilde{Q}^N = \psi_M^0(\tilde{Q}_0^N + \tilde{Q}_0^{N,-}(M - e) + \tilde{A}^N - \tilde{S}^N - \sqrt{N}(1 - \rho^N)e), \quad (6.9)$$

where  $M = \{M(t), t \geq 0\}$  is the renewal function associated with  $F$  and here we have set  $a = 0$ . This representation is used in the proof of our main result of this Section, Theorem 2.

We are now in a position to state the first main result of this Subsection which provides a joint convergence result on the diffusion scaled initial number of customers in the system at time 0– and the diffusion scaled, centered arrival and departure processes. We have the following.

**Proposition 9** *If  $\tilde{Q}_0^N \Rightarrow \tilde{Q}_0$  as  $N \rightarrow \infty$ , then  $(\tilde{Q}_0^N, \tilde{A}^N, \tilde{S}^N) \Rightarrow (\tilde{Q}_0, \tilde{\xi}, \tilde{\gamma})$  in  $(\mathbb{R} \times D^2[0, \infty), \mathcal{B}(\mathbb{R}) \times \mathcal{D}^2)$  as  $N \rightarrow \infty$ , where each of the limiting quantities above are independent of one another and  $\tilde{\gamma}$  is a centered, Gaussian process with  $\mathbb{P}$ -a.s. continuous sample paths and covariance function given by.*

$$E[\tilde{\gamma}(t)\tilde{\gamma}(t+\delta)] = 2 \int_0^t (M(u) - u + .5) du + \int_0^t \int_0^{t+\delta} M(t-a)M(t+\delta-b) dF(a+b),$$

for  $t, \delta \geq 0$ .

**Proof.** Note first from Proposition 2 and the definition of  $\tilde{S}^N(t)$  in (6.7), the representation

$$\tilde{S}^N(t) = \tilde{N}^N(t) + \tilde{S}_F^N(t) + \int_0^t \tilde{N}^N(t-s) dM(s) + \int_0^t \tilde{S}_F^N(t-s) dM(s), \quad t \geq 0, \quad (6.10)$$

where  $\tilde{N}^N(t) = N^{-1/2}N^N(t)$  and  $\tilde{S}_F^N(t) = N^{-1/2}S_F^N(t)$ .

It therefore follows by Propositions 5 and 6 that

$$\tilde{S}^N(t) = - \left( \tilde{W}_0^N(t) + \tilde{M}_2^N(t) + \int_0^t \tilde{W}_0^N(t-s) dM(s) + \int_0^t \tilde{M}_2^N(t-s) dM(s) \right), \quad (6.11)$$

for  $t \geq 0$ , where  $\tilde{W}_0^N(t) = N^{-1/2}W_0^N(t)$  and  $\tilde{M}_2^N(t) = N^{-1/2}M_2^N(t)$ .

Now let  $f : \mathbb{R} \times D^3[0, \infty) \mapsto \mathbb{R} \times D^2[0, \infty)$  be the function defined as follows. For each  $x = (x_1, x_2, x_3, x_4) \in \mathbb{R} \times D^3[0, \infty)$ , we set

$$f(x) = (f_1(x_1), f_2(x_2), f_3(x_3, x_4)), \quad (6.12)$$

where  $f_1(x_1) = x_1$ ,  $f_2(x_2) = x_2$ , and

$$f_3(x_3, x_4)(\cdot) = - \left( x_3(\cdot) + \int_0^\cdot x_3(\cdot - s) dM(s) + x_4(\cdot) + \int_0^\cdot x_4(\cdot - s) dM(s) \right). \quad (6.13)$$

It then follows (6.11) and the definition of  $f$  in (6.12) and (6.13), that we have the representation

$$(\tilde{Q}_0^N, \tilde{A}^N, \tilde{S}^N) = f(\tilde{Q}_0^N, \tilde{A}^N, \tilde{W}_0^N, \tilde{M}_2^N).$$

Next, note that since by assumption  $\tilde{Q}_0^N \Rightarrow \tilde{Q}_0$  as  $N \rightarrow \infty$ , it follows by Proposition 7 of [11] that

$$(\tilde{Q}_0^N, \tilde{W}_0^N, \tilde{A}^N, \tilde{M}_2^N) \Rightarrow (\tilde{Q}_0, \tilde{W}_0(F_e), \tilde{\xi}, \tilde{M}_2) \text{ in } (\mathbb{R} \times D^3[0, \infty), \mathcal{B}(\mathbb{R}) \times \mathcal{D}^3) \quad (6.14)$$

as  $N \rightarrow \infty$ , where each of the limiting processes above are independent of one another and  $\mathbb{P}$ -a.s. continuous. Thus, since by Lemma 1 of the Appendix we have that  $f : (\mathbb{R} \times D^3[0, \infty), |\cdot| \times d_{J_1}^3) \mapsto (\mathbb{R} \times D^2[0, \infty), |\cdot| \times d_{J_1}^2)$  is continuous at continuous limit points  $(x_1, x_2, x_3, x_4) \in \mathbb{R} \times C^3[0, \infty)$ , it follows that

$$\mathbb{P}((\tilde{Q}_0, \tilde{W}_0(F_e), \tilde{\xi}, \tilde{M}_2) \in \text{Disc}(f)) = 0, \quad (6.15)$$

and so, by the measurability of  $f : (\mathbb{R} \times D^3[0, \infty), \mathcal{B}(\mathbb{R}) \times \mathcal{D}^3) \mapsto (\mathbb{R} \times D^2[0, \infty), \mathcal{B}(\mathbb{R}) \times \mathcal{D}^2)$  given by Lemma 1 in the Appendix, we have by the Continuous Mapping Theorem 3.4.3 of [14]  $(\tilde{Q}_0^N, \tilde{A}^N, \tilde{S}^N) \Rightarrow (\tilde{Q}_0, \tilde{\xi}, \tilde{\gamma})$  in  $(\mathbb{R} \times D^2[0, \infty), \mathcal{B}(\mathbb{R}) \times \mathcal{D}^2)$  as  $N \rightarrow \infty$ , where  $\tilde{Q}_0, \tilde{\xi}$ , and  $\tilde{\gamma}$  are independent of one another and

$$\tilde{\gamma}(t) = - \left( \tilde{W}_0(F_e(t)) + \tilde{M}_2(t) + \int_0^t \tilde{W}_0(F_e(t-s)) dM(s) + \int_0^t \tilde{M}_2(t-s) dM(s) \right), \quad (6.16)$$

for  $t \geq 0$ .

It now remains to show that the covariance function of  $\tilde{\gamma}$  is given by (6.10). In order to show this, first, for each  $N \geq 1$ , define

$$\tilde{R}^N(t) = \frac{\sum_{i=1}^N (R_i(t) - t)}{\sqrt{N}}, \quad t \geq 0,$$

where  $\{R_i, i \geq 1\}$  is an i.i.d. sequence of delayed renewal processes with initial delay distribution given by  $F_e$  and subsequent interarrival distributions given by  $F$ . Also, set  $\tilde{R}^N = \{\tilde{R}^N(t), t \geq 0\}$ . It may then be shown using an approach similar to the proof of Proposition 2 and the proof given directly above that we also have the convergence  $\tilde{R}^N \Rightarrow \tilde{\gamma}$  as  $N \rightarrow \infty$ . Thus, it must be the case that the covariance function of  $\tilde{\gamma}$  is given by that of the individual  $R_i$ 's, which, by Theorem 7.2.4 of [14], is given by (6.10) above. This completes the proof.  $\square$

Now note the following regarding Proposition 9. First, by Proposition 9 it follows that  $\tilde{S}^N \Rightarrow \tilde{\gamma}$  as  $N \rightarrow \infty$ , where the covariance function of  $\tilde{\gamma}$  is given by (6.10). Furthermore, for each  $N \geq 1$ , setting

$$\tilde{R}^N(t) = \frac{\sum_{i=1}^N (R_i(t) - t)}{\sqrt{N}}, \quad t \geq 0,$$

where  $\{R_i, i \geq 1\}$  is an i.i.d. sequence of delayed renewal processes with initial delay distribution given by  $F_e$  and subsequent interarrival distributions given by  $F$  and letting  $\tilde{R}^N = \{\tilde{R}^N(t), t \geq 0\}$ , we have that  $\tilde{R}^N \Rightarrow \tilde{\gamma}$  as well as  $N \rightarrow \infty$ . However, for each  $i = 1, \dots, \min(Q_0, N)$ , we have that the processes  $S_i$  are also i.i.d renewal processes with initial residual distribution given by  $F_e$  and subsequent interarrival times given by  $F$ . Thus, it follows by the definition of  $\hat{S}_i$  in (2.3) and (2.5) and of  $\hat{S}$  and  $\tilde{S}$  in (2.7) and (6.7) that one of the results of Proposition 9 is that asymptotically in the limit, as far as the sequence of centered process  $\{\tilde{S}^N, N \geq 1\}$  is concerned, the idle time of each individual server becomes negligible.

We are now in a position to present our main result of this Subsection, Theorem 2, which provides a limit for the sequence of diffusion scaled queue length processes,  $\{\tilde{Q}^N, N \geq 1\}$ . The interested reader may check to see that the limiting process obtained here is the same as that in Corollary 3 of [11]. We have the following.

**Theorem 2** *If the residual service time  $F_0 = F_e$  and  $\tilde{Q}_0^N \Rightarrow \tilde{Q}_0$  as  $N \rightarrow \infty$ , then  $\tilde{Q}^N \Rightarrow \psi_M^0(\tilde{Q}_0 + \tilde{Q}_0^-(M - e) + \tilde{\xi} - \tilde{\gamma} - \beta e)$  as  $N \rightarrow \infty$ .*

**Proof.** The proof is similar in spirit to the proof of Theorem 3 of [11]. First note that since  $\tilde{Q}_0^N \Rightarrow \tilde{Q}_0$  as  $N \rightarrow \infty$ , it follows by Proposition 9 that

$$(\tilde{Q}_0^N, \tilde{A}^N, \tilde{S}^N) \Rightarrow (\tilde{Q}_0, \tilde{\xi}, \tilde{\gamma}) \text{ in } (\mathbb{R} \times D^2[0, \infty), |\cdot|) \times d_{J_1}^2 \text{ as } N \rightarrow \infty,$$

where each of the limiting processes appearing on the righthand side above are independent of one another. Next, since the metric spaces  $(\mathbb{R}, |\cdot|)$  and  $(D[0, \infty), d_{J_1})$  are both separable, it follows by Theorem 11.4.1 of [14] that the product space  $(\mathbb{R} \times D^2[0, \infty), |\cdot|) \times d_{J_1}^2$  is separable as well and hence, by the Skorohod Representation Theorem [14], there exists an alternative probability space  $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$  on which are defined a sequence of processes  $\{(\hat{Q}_0^N, \hat{A}^N, \hat{S}^N), N \geq 1\}$  such that

$$(\hat{Q}_0^N, \hat{A}^N, \hat{S}^N) \stackrel{d}{=} (\tilde{Q}_0^N, \tilde{A}^N, \tilde{S}^N), \quad N \geq 1, \quad (6.17)$$

and a single process  $(\hat{Q}_0, \hat{\xi}, \hat{\gamma}) \stackrel{d}{=} (\tilde{Q}_0, \xi, \gamma)$  such that

$$(\hat{Q}_0^N, \hat{A}^N, \hat{S}^N) \rightarrow (\hat{Q}_0, \hat{\xi}, \hat{\gamma}) \text{ in } (\mathbb{R} \times D^2[0, \infty), |\cdot| \times d_{J_1}^2) \hat{\mathbb{P}}\text{-a.s. as } N \rightarrow \infty. \quad (6.18)$$

Also note that since the processes  $\hat{\xi}$  and  $\hat{\gamma}$  are  $\hat{\mathbb{P}}$ -a.s. continuous, it follows that the above convergence may be strengthened to convergence in  $(\mathbb{R} \times D^2[0, \infty), |\cdot| \times u^2)$ .

Now note that by (6.17) we have that

$$\hat{Q}_0^{N,-}(M - e) - \sqrt{N}(1 - \rho^N)e \stackrel{d}{=} \tilde{Q}_0^{N,-}(M - e) - \sqrt{N}(1 - \rho^N)e, \quad N \geq 1.$$

Furthermore, by (6.3) and (6.18), it follows that for each  $T \geq 0$ ,

$$\begin{aligned} & \sup_{0 \leq t \leq T} |(\hat{Q}_0^{N,-}(M(t) - t) - \sqrt{N}(1 - \rho^N)t) - (\hat{Q}_0^-(M(t) - t) - \beta t)| \\ &= \sup_{0 \leq t \leq T} |(\hat{Q}_0^{N,-} - \hat{Q}_0^-)(M(t) - t) - (\sqrt{N}(1 - \rho^N) - \beta)t| \\ &\leq \sup_{0 \leq t \leq T} |(\hat{Q}_0^{N,-} - \hat{Q}_0^-)(M(t) - t)| + \sup_{0 \leq t \leq T} |(\sqrt{N}(1 - \rho^N) - \beta)t| \\ &\leq (T + M(T))|\hat{Q}_0^{N,-} - \hat{Q}_0^-| + T|\sqrt{N}(1 - \rho^N) - \beta| \\ &\rightarrow 0 \hat{\mathbb{P}}\text{-a.s. as } N \rightarrow \infty, \end{aligned}$$

so that

$$\hat{Q}_0^{N,-}(M - e) - \sqrt{N}(1 - \rho^N)e \rightarrow \hat{Q}_0^-(M - e) - \beta e \text{ in } (D[0, \infty), u) \hat{\mathbb{P}}\text{-a.s. as } N \rightarrow \infty. \quad (6.19)$$

Now, for each  $N \geq 1$ , set

$$\hat{Q}^N = \psi_M^0(\hat{Q}_0^N + \hat{Q}_0^{N,-}(M - e) + \hat{A}^N - \hat{S}^N - \sqrt{N}(1 - \rho^N)e) \quad (6.20)$$

and note that since by (6.17),

$$\begin{aligned} & \hat{Q}_0^N + \hat{Q}_0^{N,-}(M - e) + \hat{A}^N - \hat{S}^N - \sqrt{N}(1 - \rho^N)e \\ &\stackrel{d}{=} \tilde{Q}_0^N + \tilde{Q}_0^{N,-}(M - e) + \tilde{A}^N - \tilde{S}^N - \sqrt{N}(1 - \rho^N)e, \end{aligned}$$

it follows by the representation (6.9) and the measurability of  $\psi_M^0$  from Proposition 7 that  $\hat{Q}^N \stackrel{d}{=} \tilde{Q}^N$ . Furthermore, by (6.18) and (6.19), we have that

$$\begin{aligned} & \hat{Q}_0^N + \hat{Q}_0^{N,-}(M - e) + \hat{A}^N - \hat{S}^N - \sqrt{N}(1 - \rho^N)e \\ &\rightarrow \hat{Q}_0^+ \hat{Q}_0^-(M - e) + \hat{\xi} - \hat{\gamma} - \beta e, \text{ in } (D[0, \infty), u) \hat{\mathbb{P}}\text{-a.s. as } N \rightarrow \infty, \end{aligned}$$

and so, by (6.20) and the continuity portion of Proposition 7, it follows that

$$\hat{Q}^N \rightarrow \psi_M^0(\hat{Q}_0 + \hat{Q}_0^-(M - e) + \hat{\xi} - \hat{\gamma} - \beta e) \hat{\mathbb{P}} \text{ a.s. as } N \rightarrow \infty.$$

However, since  $(\hat{Q}_0, \hat{\xi}, \hat{\gamma}) \stackrel{d}{=} (\tilde{Q}_0, \xi, \gamma)$ , it follows by the measurability portion of Proposition 7 that

$$\psi_M^0(\hat{Q}_0 + \hat{Q}_0^-(M - e) + \hat{\xi} - \hat{\gamma} - \beta e) \stackrel{d}{=} \psi_M^0(\tilde{Q}_0 + \tilde{Q}_0^-(M - e) + \tilde{\xi} - \tilde{\gamma} - \beta e),$$

which completes the proof.  $\square$

Note that by the definition of the regular map  $\Psi_M^0$  given by (4.1), we have that the limit of Theorem 2 may be represented as  $\tilde{Q}_M$ , the unique strong solution to the stochastic convolution equation

$$\tilde{Q}_M(t) = \tilde{Q}_0 + \tilde{Q}_0^-(M(t) - t) + \tilde{\xi}(t) - \tilde{\gamma}(t) - \beta t - \int_0^t \tilde{Q}^-(t - s) dM(s), \quad t \geq 0. \quad (6.21)$$

We now attempt provide a term by term explanation of the quantities appearing on the righthand side (6.21). First, the quantity  $\tilde{Q}_0$  represents the initial number of customers in the system at time 0-. Next, for each  $t \geq 0$ , the term  $\tilde{\xi}(t)$  is representative of the randomness arising from the number of arrivals which have occurred to the system by time  $t$ , while the term  $\tilde{\gamma}(t)$  is representative of the randomness arising from the number of departures from the system by time  $t$ . The term  $\tilde{Q}_0^-(M(t) - t)$  is a slight adjustment necessary to take into account those servers which were idle at time 0- and while the term  $\beta t$  is representative of the capacity imbalance between the arrival rate and service rate of the system. Finally, note that by the discussion following the proof of Proposition 9 it follows that the term  $\tilde{\gamma}(t)$  does not take into account the idleness of the servers in the system. Thus, the final integral term appearing on the righthand side of (6.21) is necessary to modify  $\tilde{\gamma}(t)$  in order to keep track of the idleness of the servers in the system.

In Subsection 6.5, we show that the limit  $\tilde{Q}_M$  given above may be equivalently represented as  $\tilde{Q}_F$ , the limit given by Theorem 2 of [11]. However, in the following Subsection, we concentrate on studying the amount of time that each server is allowed to wait between serving customers.

## 6.4 The Virtual Idle Time Process

In this Subsection, we study the amount of time that each server is allowed to wait between serving customers. Recall that we are assuming that the servers are operating under the longest-idle-first-served routing policy although this assumption has not been necessary for our results obtained thus far.

For each  $N \geq 1$ , let  $I^N(t)$  denote the amount of time that a hypothetical server becoming free at time  $t \geq 0$  would have to wait before receiving its next customer and, for each  $k \geq 1$ , let  $I_k^N$  denote the amount of time that the  $k^{\text{th}}$  server to become free after time 0- has to wait before receiving its next customer. In this Subsection, we study the diffusion scaled virtual idle time process  $\tilde{I}^N = \{N^{1/2}I^N(t), t \geq 0\}$  and the diffusion scaled server waiting time process  $\tilde{I}^N = \{N^{1/2}I_{[Nt]}^N, t \geq 0\}$  for the  $G/GI/N$  queue in the Halfin-Whitt. Our main result is to show that each of these processes converges to an identical limiting process as  $N$  goes to  $\infty$ , which may be explicitly characterized.

Note that as a byproduct this result also implies that the amount of time that a typical server is allowed to idle between the service of consecutive customers is on the order of  $N^{-1/2}$  of the number of servers in the system. The following is now the main result of this Subsection.

**Proposition 10** *If the residual service time distribution  $F_0 = F_e$  and  $\tilde{Q}_0^N \Rightarrow \tilde{Q}_0$  as  $N \rightarrow \infty$ , then  $\tilde{I}^N \Rightarrow \tilde{I}$  as  $N \rightarrow \infty$  and  $\hat{I}^N \Rightarrow \tilde{I}$  as  $N \rightarrow \infty$ , where  $\tilde{I}$  is given by the unique strong solution to the integral equation*

$$\tilde{I}(t) = \left( -\tilde{Q}_0 - \tilde{Q}_0^-(M - e) - \tilde{\xi} + \tilde{\gamma} + \beta e - \int_0^t \tilde{I}(t-s) dM(s) \right)^+, \quad t \geq 0.$$

**Proof.** Recall first the definition of virtual idle time,  $I^N(t)$ , as the amount of time that a hypothetical server becoming free in the  $N^{\text{th}}$  system at time  $t \geq 0$  would have to wait before receiving its next customer. Our first step is represent  $I^N(t)$  in terms of a first passage time process.

We begin by noting that for each  $t \geq 0$  we have that the quantity  $(Q_0^N - N)^+ + A^N(t) - (Q^N(t) - N)^+$  represents the number of customers who were not in service at time 0– but who have entered service by time  $t$ . Next, since servers are assigned customers on a longest-idle-first-served basis, it follows that a server becoming idle at time  $t$  would have to remain idle until at least the smallest time  $u \geq t$  such that the number of customers who have entered service by time  $u$  is at least equal to the number of idle servers at time 0– plus the number of departures by time  $t$ . Thus, we have that

$$I^N(t) + t = \inf\{u \geq 0 : (Q_0^N - N)^+ + A^N(u) - (Q^N(u) - N)^+ > (N - Q_0^N)^+ + S^N(t)\} \quad (6.22)$$

for  $t \geq 0$ .

Now note that by Proposition 9,  $(\tilde{Q}_0^N, \tilde{A}^N, \tilde{S}^N) \Rightarrow (\tilde{Q}_0, \tilde{\xi}, \tilde{\gamma})$  in  $(\mathbb{R} \times D^2[0, \infty), |\cdot| \times d_{J_1}^2)$  as  $N \rightarrow \infty$  and so, by the definition of  $\tilde{A}^N$  in (6.1) and (6.3), we have, setting

$$\check{A}^N(t) = \frac{A^N(t) - Nt}{\sqrt{N}}, \quad t \geq 0,$$

and  $\check{A}^N = \{\check{A}^N(t), t \geq 0\}$ , that  $(\tilde{Q}_0^N, \check{A}^N) \Rightarrow (\tilde{Q}_0, \tilde{\xi} + \beta e)$  in  $(\mathbb{R} \times D[0, \infty), |\cdot| \times d_{J_1})$  as  $N \rightarrow \infty$ . Furthermore, by Proposition 8 of [11],  $(\tilde{Q}^N, \tilde{Q}^{N,+}) \Rightarrow (\tilde{Q}, \tilde{Q}^+)$  as  $N \rightarrow \infty$ . It therefore follows by Prohorov's Theorem [1] that both of the sequences  $\{(\tilde{Q}_0^N, \check{A}^N), N \geq 1\}$  and  $\{(\tilde{Q}^N, \tilde{Q}^{N,+}), N \geq 1\}$  are tight. Thus, the sequence  $\{(\tilde{Q}_0^N, \check{A}^N, \tilde{Q}^N, \tilde{Q}^{N,+}), N \geq 1\}$  is tight in  $(\mathbb{R} \times D^3[0, \infty), |\cdot| \times d_{J_1}^4)$  as well and so, by Prohorov's Theorem [1], it is relatively compact. However, by the definition of relative compactness, this then implies that for every sequence  $\{N_k\}$ , there exists a further subsequence,  $\{N'_k\}$ , such that

$$(\tilde{Q}_0^{N'_k}, \check{A}^{N'_k}, \tilde{Q}^{N'_k}, \tilde{Q}^{N'_k,+}) \Rightarrow (\hat{Q}_0, \hat{\xi}, \hat{Q}, \hat{Q}^+) \quad \text{as } k \rightarrow \infty. \quad (6.23)$$

Now note that for each  $N \geq 1$  and  $t \geq 0$ , setting

$$\begin{aligned}\tilde{X}^N(t) &= N^{-1/2}((Q_0^N - N)^+ + A^N(t) - (Q^N(t) - N)^+ - Nt) \\ &= \tilde{Q}_0^{N,+} + \check{A}^N(t) - \tilde{Q}^{N,+}(t),\end{aligned}$$

and

$$\begin{aligned}\tilde{Y}^N(t) &= N^{-1/2}((N - Q_0^N)^+ + S^N - Nt) \\ &= N^{-1/2}((N - Q_0^N)^+ + Q_0^N + A^N(0) - Q^N(t) - Nt) \\ &= N^{-1/2}(-(Q_0^N - N)^- + (Q_0^N - N) + (A^N(t) - Nt) - (Q^N(t) - N)) \\ &= N^{-1/2}((Q_0^N - N)^+ + (A^N(t) - Nt) - (Q^N(t) - N)) \\ &= \tilde{Q}_0^{N,+} + \check{A}^N(t) - \tilde{Q}^N(t),\end{aligned}$$

and letting  $\tilde{X}^N = \{\tilde{X}^N(t), t \geq 0\}$  and  $\tilde{Y}^N = \{\tilde{Y}^N(t), t \geq 0\}$ , it follows by the Continuous Mapping Theorem [14] and (6.23) that

$$(\tilde{X}^{N'_k}, \tilde{Y}^{N'_k}) \Rightarrow (\hat{Q}_0^+ + \hat{\xi} + \beta e - \hat{Q}^+, \hat{Q}_0 + \hat{\xi} + \beta e - \hat{Q}) \text{ as } k \rightarrow \infty.$$

Thus, by the Corollary in [10], we that as  $k \rightarrow \infty$ ,

$$\tilde{I}^{N'_k} \Rightarrow (\hat{Q}_0 + \hat{\xi} + \beta e - \hat{Q}) - (\hat{Q}_0^+ + \hat{\xi} + \beta e - \hat{Q}^+) = -\hat{Q} + \hat{Q}^+ = -\hat{Q}^-.$$

However, since  $\hat{Q} \stackrel{d}{=} \tilde{Q}$  and thus  $\hat{Q}^- \stackrel{d}{=} \tilde{Q}^-$  as well and, since the sequence  $\{N_k\}$  was arbitrary, it follows that  $\tilde{I}^N \Rightarrow -\tilde{Q}^-$  as  $N \rightarrow \infty$ . However, since by Theorem 2 we have that

$$-\tilde{Q}^-(t) = \left( -\tilde{Q}_0 - \tilde{Q}_0^-(M - e) - \tilde{\xi} + \tilde{\gamma} + \beta e - \int_0^t -\tilde{Q}^-(t-s) dM(s) \right)^-, \quad t \geq 0,$$

this completes the proof of the first part of the Proposition.

For the second part of the Proposition, first note that recalling from (3.1) the definition of

$$d_k^N = \inf\{t \geq 0 : S^N(t) \geq k\}, \quad k \geq 1, \quad (6.24)$$

as the time of the  $k^{\text{th}}$  departure from the  $N^{\text{th}}$  system, we have, by definition, that  $I_k^N = I^N(d_k^N)$  and so it follows that  $\tilde{I}^N = \{N^{1/2}I^N(d_{[Nt]}^N), t \geq 0\}$ . Next, setting  $\check{S}^N = \{N^{-1}S^N(t), t \geq 0\}$ , we claim that  $\check{S}^N \Rightarrow e$  as  $N \rightarrow \infty$ . This follows in a straightforward manner since by the assumption of the Proposition and Corollary 1, as  $N \rightarrow \infty$ ,

$$\check{S}^N = \bar{Q}_0^N + \bar{A}^N - \bar{Q}^N \Rightarrow 1 + e - 1 = e.$$

By (6.24) above it now follows that  $d_{[N\cdot]}^N \Rightarrow e$  as  $N \rightarrow \infty$  and so, since  $e$  is a deterministic process, we have by Theorem 3.9 of [1] that  $(\tilde{I}^N, d_{[N\cdot]}^N) \Rightarrow (\tilde{I}, e)$  in  $(D^2[0, \infty), d_{J_1}^2)$  as  $N \rightarrow \infty$ . Thus, by the Random Time Change Theorem [1], it follows that  $\tilde{I}^N = \tilde{I}^N \circ d_{[N\cdot]}^N \Rightarrow \tilde{I} \circ e = \tilde{I}$  as  $N \rightarrow \infty$ , which completes the proof.  $\square$

## 6.5 An Alternative Representation

We now complete this Section by showing that  $\tilde{Q}_M$ , the limit of the sequence of diffusion scaled processes obtained in Theorem 2 of the present paper, is equivalent to  $\tilde{Q}_F$ , the limit given by Theorem 2 of [11]. For each  $t \geq 0$ , let

$$\tilde{M}_Q(t) = \tilde{Q}_0^+(G(t) - \bar{F}_e(t)) \quad (6.25)$$

and set

$$\tilde{Q}_I(t) = \tilde{Q}_0 \bar{F}_e(t) + \tilde{W}_0(F_e(t)) + \int_0^t G(t-s) \tilde{\xi}(s) + \tilde{M}_2(t), \quad (6.26)$$

where  $\tilde{W}_0 = \{\tilde{W}_0(t), t \geq 0\}$  is a Brownian bridge,  $\tilde{M}_2 = \{\tilde{M}_2(t), t \geq 0\}$  is a centered, Gaussian process whose covariance function may be explicitly calculated (see Proposition 5 of [11]) and each of the four processes appearing in (6.26) above are assumed to be independent of one another. Note also that the integral with respect to  $\tilde{\xi}$  in (6.26) is meant to be interpreted as the result of integration by parts. For further information on the above defined processes, the interested reader is also referred to Section 5 of the prequel [11].

We may now state the following result, which is the main result of this Subsection.

**Corollary 2** *The limiting process,  $\tilde{Q}_M$ , of Theorem 2 may be equivalently represented as the unique strong solution to*

$$\tilde{Q}_F(t) = \tilde{M}_Q(t) + \tilde{Q}_I(t) - \beta F_e(t) + \int_0^t \tilde{Q}_F^+(t-s) dF(s), \quad \text{for } t \geq 0, \quad (6.27)$$

where  $\tilde{Q}_F^+ = \max(\tilde{Q}_F, 0)$ .

**Proof.** First note since we have the decomposition  $\tilde{Q}_M = \tilde{Q}_M^+ + \tilde{Q}_M^-$ , it follows by (6.21) above that

$$\tilde{Q}_M^-(t) = -\tilde{Q}_M^+(t) + \tilde{Q}_0 + \tilde{Q}_0^-(M(t) - t) + \tilde{\xi}(t) - \tilde{\gamma}(t) - \beta t - \int_0^t \tilde{Q}_M^-(t-s) dM(s), \quad (6.28)$$

for  $t \geq 0$ .

Next, note that (6.28) is an integral equation of renewal type in terms of  $\tilde{Q}_M^- = \{\tilde{Q}_M^-(t), t \geq 0\}$ . Furthermore, since  $-\tilde{Q}_M^+ + \tilde{Q}_0 + \tilde{Q}_0^-(M - e) + \tilde{\xi} - \tilde{\gamma} - \beta e$  is  $\mathbb{P}$ -a.s. a locally bounded function, being an element of  $D[0, \infty)$ , it follows that the unique solution to (6.28) is given by

$$\begin{aligned} \tilde{Q}_M^-(t) &= -\tilde{Q}_M^+(t) + \tilde{Q}_0 + \tilde{Q}_0^-(M(t) - t) + \tilde{\xi}(t) - \tilde{\gamma}(t) - \beta t \\ &\quad + \int_0^t \tilde{Q}_M^+(t-s) dF(s) - \tilde{Q}_0 F(t) - \tilde{Q}_0^- \int_0^t (M(t-s) - (t-s)F(s)) \\ &\quad - \int_0^t \tilde{\xi}(t-s) dF(s) + \int_0^t \tilde{\gamma}(t-s) F(s) + \beta \int_0^t (t-s) dF(s), \end{aligned} \quad (6.29)$$

for  $t \geq 0$ . Moreover, noting that integrating by parts we have,

$$\int_0^t (t-s)dF(s) = -\int_0^t F(s)ds = -t + F_e(t),$$

for  $t \geq 0$ , and recalling from (2.2) the renewal equation

$$M(t) = F(t) + \int_0^t F(t-s)dM(s), \quad t \geq 0, \quad (6.30)$$

it follows from (6.29) that, after a little bit of algebra,

$$\begin{aligned} \tilde{Q}_M(t) &= \tilde{Q}_0 \bar{F}_e(t) + \tilde{Q}_0^+(G(t) - \bar{F}_e(t)) + \tilde{\xi}(t) + \int_0^t \tilde{\xi}(t-s)dG(s) - \tilde{\gamma}(t) + \int_0^t \tilde{\gamma}(t-s)F(s) \\ &\quad - \beta F_e(t) + \int_0^t \tilde{Q}_M^+(t-s)dF(s), \end{aligned} \quad (6.31)$$

for  $t \geq 0$ . Furthermore, since by (6.16) in the proof of Proposition 9 we have that

$$\tilde{\gamma}(t) = -\left( \tilde{W}_0(F_e(t)) + \tilde{M}_2(t) + \int_0^t \tilde{W}_0(F_e(t-s))dM(s) + \int_0^t \tilde{M}_2(t-s)dM(s) \right),$$

for  $t \geq 0$ , it follows by the renewal equation (6.30) above that

$$\begin{aligned} &-\tilde{\gamma}(t) + \int_0^t \tilde{\gamma}(t-s)F(s) \quad (6.32) \\ &= \tilde{W}_0(F_e(t)) + \tilde{M}_2(t) + \int_0^t \tilde{W}_0(F_e(t-s))dM(s) + \int_0^t \tilde{M}_2(t-s)dM(s) \\ &\quad - \int_0^t \left( \tilde{W}_0(F_e(t-s)) + \tilde{M}_2(t-s) \right) dF(s) \\ &\quad - \int_0^t \left( \int_0^{t-s} \tilde{W}_0(F_e(t-s-u))dM(u) + \int_0^{t-s} \tilde{M}_2(t-s-u)dM(u) \right) dF(u) \\ &= \tilde{W}_0(F_e(t)) - \int_0^t \tilde{W}_0(F_e(t-s))d \left( M(s) - dF(s) - \int_0^s F(s-u)dM(u) \right) \\ &\quad + \tilde{M}_2(t) - \int_0^t \tilde{M}_2(t)d \left( M(s) - dF(s) - \int_0^s F(s-u)dM(u) \right) \\ &= \tilde{W}_0(F_e(t)) + \tilde{M}_2(t). \end{aligned}$$

Thus, substituting (6.32) into (6.31) and noting that integrating by parts we obtain

$$\tilde{\xi}(t) + \int_0^t \tilde{\xi}(t-s)dG(s) = \int_0^t G(t-s)d\tilde{\xi}(s), \quad t \geq 0, \quad (6.33)$$

it follows that

$$\begin{aligned}\tilde{Q}_M(t) &= \tilde{Q}_0^+(G(t) - \bar{F}_e(t)) + \tilde{Q}_0\bar{F}_e(t) + \tilde{W}_0(F_e(t)) + \int_0^t G(t-s)\tilde{\xi}(s) + \tilde{M}_2(t) \\ &\quad - \beta F_e(t) + \int_0^t \tilde{Q}_M^+(t-s)dF(s),\end{aligned}$$

for  $t \geq 0$ , which, by (6.25) and (6.26), now completes the proof.  $\square$

## 7 Conclusion

In this paper, we studied the  $G/GI/N$  queue in the Halfin-Whitt regime from the point of view of the servers in the system. We have labeled this approach the “Idle Time System Equations” approach and note that it stands opposite to the “Infinite Server Queue System Equations” approach first considered in the prequel [11]. Our first main result of the current paper, Theorem 1, provides a limit for the fluid scaled queue length process of the  $G/GI/N$  queue in the Halfin-Whitt regime. Next, in Theorem 2, we provide a limit for the diffusion scaled and properly centered queue length process. This limit may best be summarized as the solution to a stochastic convolution equation. In Corollary 2, we show how the limit obtained in Theorem 2 of [11] using the “Infinite Server Queue System Equations” may be derived as a consequence of Theorem 2 of the present paper.

In the future, depending upon the particular application, one may find it more or less useful to use the “Infinite Server Queue System Equations” approach as opposed to the “Idle Time System Equations Approach”. In future work, we intend to illustrate just how such situations may arise.

## 8 Appendix

In the Appendix, we provide the proofs of Proposition 7 and Lemma 1. We begin with the proof of Proposition 7.

**Proof of Proposition 7.** Recall first the definition of  $\varphi_B^a(x)$  from (4.2) as the unique solution  $z \in D[0, \infty)$  to

$$z(t) = x(t) + \int_0^t (z(t-s) + a)^+ dB(s), \quad t \geq 0, \quad (8.1)$$

where  $a^+ = \max(0, a)$ , and also the definition of  $\vartheta_B^a : D[0, \infty) \mapsto D[0, \infty)$  from (4.3) as given by

$$\vartheta_B^a(x)(t) = x(t) - \int_0^t x(t-s)dB(s) + aB(t), \quad t \geq 0. \quad (8.2)$$

Our proofs below will constructively show that  $\Psi_R^a = \varphi_B^a \circ \vartheta_B^a$ , where  $B$  is the distribution function associated with renewal function  $R$ .

**Existence:** Letting  $z = \varphi_B^a \circ \vartheta_B^a(x)$ , where  $B$  is the distribution function associated the renewal function  $R$ , it follows by the definition of  $\varphi_B^a$  in (8.1) and the definition of  $\vartheta_B^a$  in (8.2) that

$$z(t) = x(t) - \int_0^t x(t-s)dB(s) + aB(t) + \int_0^t (z(t-s) + a)^+ dB(s), \quad t \geq 0. \quad (8.3)$$

Thus, noting the decomposition  $z = a + (z+a)^+ + (z+a)^-$ , we have that

$$\begin{aligned} (z(t) + a)^+ &= -a - (z(t) + a)^- + x(t) - \int_0^t x(t-s)dB(s) + aB(t) \\ &\quad + \int_0^t (z(t-s) + a)^+ dB(s), \end{aligned} \quad (8.4)$$

for  $t \geq 0$ . Now note that (8.4) is an integral equation of renewal type (in terms of  $(z+a)^+$ ), whose unique solution [8] is given by

$$\begin{aligned} &(z(t) + a)^+ \\ &= -a - (z(t) + a)^- + x(t) - \int_0^t x(t-s)dB(s) + aB(t) \\ &\quad + \int_0^t \left( -a - (z(t-s) + a)^- + x(t-s) - \int_0^{t-s} x(t-s-u)dB(u) + aB(t-s) \right) dR(s), \end{aligned} \quad (8.5)$$

for  $t \geq 0$ . However, since by the renewal equation (2.2) we have that

$$\begin{aligned} & - \int_0^t x(t-s)dB(s) + \int_0^t x(t-s)dR(s) - \int_0^t \left( \int_0^{t-s} x(t-s-u)dB(u) \right) dR(s) \\ &= \int_0^t x(t-s)d \left( -B(s) + R(s) - \int_0^s B(s-u)R(u) \right) \\ &= 0, \end{aligned}$$

and, similarly,

$$aB(t) - aR(t) + a \int_0^t B(t-s)dR(s) = 0,$$

it follows from (8.5) that

$$(z(t) + a)^+ = -a - (z(t) + a)^- + x(t) - \int_0^t (z(t-s) + a)^- dR(s), \quad t \geq 0,$$

which is equivalent to

$$z(t) = x(t) - \int_0^t (z(t-s) + a)^- dR(s), \quad t \geq 0,$$

which completes the proof of existence.  $\square$

**Uniqueness:** We use the reverse argument of the existence proof above. Suppose first that  $z \in D[0, \infty)$  satisfies (4.1). It then follows by the decomposition  $z = a + (z + a)^+ + (z + a)^-$ , that

$$(z(t) + a)^- = -a - (z(t) + a)^+ + x(t) - \int_0^t (z(t-s) + a)^- dR(s), \quad t \geq 0. \quad (8.6)$$

Next, note that (8.6) is an integral equation of renewal type (in terms of  $(z + a)^-$ ), whose unique solution is given by

$$(z(t) + a)^- = -a - (z(t) + a)^+ + x(t) - \int_0^t (-a - (z(t-s) + a)^+ + x(t-s)) dB(s), \quad (8.7)$$

for  $t \geq 0$ . Next, adding  $a + (z(t) + a)^+$  to both the right and lefthand sides of (8.7), we then obtain that

$$z(t) = x(t) - \int_0^t x(t-s) dB(s) + aB(t) + \int_0^t (z(t-s) + a)^+ dB(s), \quad t \geq 0, \quad (8.8)$$

and so it now follows by Proposition 2 of [11] that the unique solution of (8.8) is given by

$$z = \varphi_B^a \left( x(\cdot) - \int_0^\cdot x(\cdot - s) dB(s) + aB(\cdot) \right), \quad (8.9)$$

which completes the proof.  $\square$

**Lipschitz Continuity:** Recall first by Proposition 2 of [11] that  $\varphi_B^a : D[0, \infty) \mapsto D[0, \infty)$  is a Lipschitz continuous function in the topology of uniform convergence over bounded intervals. Thus, since compositions of Lipschitz continuous functions are again Lipschitz continuous, it follows by the representation  $\Psi_R^a = \varphi_B^a \circ \vartheta_B^a$  given in the existence and uniqueness portions of the proof above that it remains to show that  $\vartheta_B^a$  is Lipschitz continuous in order to complete the proof.

Let  $T \geq 0$  and  $x_1, x_2 \in D[0, \infty)$ . We then have that

$$\begin{aligned} \sup_{0 \leq t \leq T} |\vartheta_B^a(x_1)(t) - \vartheta_B^a(x_2)(t)| &= \sup_{0 \leq t \leq T} |(x_1(t) - \int_0^t x_1(t-s) dB(s)) - (x_2(t) - \int_0^t x_2(t-s) dB(s))| \\ &= \sup_{0 \leq t \leq T} |(x_1(t) - x_2(t)) - \int_0^t (x_1(t-s) - x_2(t-s)) dB(s)| \end{aligned}$$

$$\begin{aligned}
&\leq \sup_{0 \leq t \leq T} |x_1(t) - x_2(t)| + \sup_{0 \leq t \leq T} \left| \int_0^t (x_1(t-s) - x_2(t-s)) dB(s) \right| \\
&\leq \sup_{0 \leq t \leq T} |x_1(t) - x_2(t)| + B(T) \sup_{0 \leq t \leq T} |x_1(t) - x_2(t)| \\
&= (1 + B(T)) \sup_{0 \leq t \leq T} |x_1(t) - x_2(t)| \\
&\leq 2 \sup_{0 \leq t \leq T} |x_1(t) - x_2(t)|,
\end{aligned}$$

which shows that  $\vartheta_B^a$  is a Lipschitz continuous function with Lipschitz constant 2, which completes the proof.  $\square$

**Measurability:** Since the composition of two measurable functions is a measurable function itself, and  $\varphi_B^a$  is measurable with respect to the Skorohod  $J_1$  topology by Proposition 2 of [11], it remains to show that  $\vartheta_B^a$  is measurable with respect to the Skorohod  $J_1$  topology as well in order to complete the proof. However, note that

$$\vartheta_B^a(x) = e(x) - \Psi_B^a(x) + aB, \quad (8.10)$$

where  $e(x) = x$  is the identity function on  $D[0, \infty)$  and

$$\Psi_B^a(x)(t) = \int_0^t (x(t-s) + a)^+ dB(s), \quad t \geq 0,$$

is as defined in the measurability portion of the proof of Proposition 2 of [11]. The identity function  $e$  is clearly measurable and  $\Psi_B^a$  was shown in the measurability portion of the proof of Proposition 2 of [11] to be measurable as well. Thus, since the sum of measurable functions is a measurable function and  $aB$  is a constant, it now follows by (8.10) that  $\vartheta_B^a$  is measurable with respect to the Skorohod  $J_1$  topology. This completes the proof.  $\square$

We now present the proof of Lemma 1. Recall first the definition of  $f : \mathbb{R} \times D^3[0, \infty) \mapsto \mathbb{R} \times D^2[0, \infty)$  in (6.12) and (6.13) as

$$f((x_1, x_2, x_3, x_4)) = (f_1(x_1), f_2(x_2), f_3(x_3, x_4)), \quad (8.11)$$

for  $(x_1, x_2, x_3, x_4) \in \mathbb{R} \times D^3[0, \infty)$ , where  $f_1(x_1) = x_1$ ,  $f_2(x_2) = x_2$ , and

$$-f_3(x_3, x_4)(\cdot) = x_3(\cdot) + \int_0^\cdot x_3(\cdot - s) dM(s) + x_4(\cdot) + \int_0^\cdot x_4(\cdot - s) dM(s). \quad (8.12)$$

We then have the following result.

**Lemma 1** *The function  $f$  defined by (8.11) and (8.12) is measurable as a map from  $(\mathbb{R} \times D^3[0, \infty), \mathcal{B}(\mathbb{R}) \times \mathcal{D}^3)$  to  $(\mathbb{R} \times D^2[0, \infty), \mathcal{B}(\mathbb{R}) \times \mathcal{D}^2)$ . Furthermore, it is continuous at continuous limits points  $(x_1, x_2, x_3, x_4) \in \mathbb{R} \times D^3[0, \infty)$  such that  $x_2, x_3, x_4 \in C[0, \infty)$*

**Proof.** We first show that the function  $f : (\mathbb{R} \times D^3[0, \infty), \mathcal{B}(\mathbb{R}) \times \mathcal{D}^3) \mapsto (\mathbb{R} \times D^2[0, \infty), \mathcal{B}(\mathbb{R}) \times \mathcal{D}^2)$  is measurable. It is clear that both  $f_1 : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$  and  $f_2 : (D[0, \infty), \mathcal{D}) \mapsto (D[0, \infty), \mathcal{D})$  are measurable since both of these functions are the identity functions. Therefore, if we may now show that  $f_3 : (D^2[0, \infty), \mathcal{D}^2) \mapsto (D[0, \infty), \mathcal{D})$  is measurable, then, by (8.11), we will have shown the measurability of  $f$ .

In order to show that  $f_3 : (D^2[0, \infty), \mathcal{D}^2) \mapsto (D[0, \infty), \mathcal{D})$  is measurable, first note that by (8.12) we have

$$f_3(x_3, x_4) = a(x_3, x_4) + b(x_3, x_4) + c(x_3, x_4) + d(x_3, x_4),$$

where  $a(x_3, x_4) = x_3, c(x_3, x_4) = x_4$ ,

$$b(x_3, x_4)(\cdot) = \int_0^\cdot x_3(\cdot - s) dM(s)$$

and

$$d(x_3, x_4)(\cdot) = \int_0^\cdot x_4(\cdot - s) dM(s).$$

The functions  $a(x_3, x_4)$  and  $c(x_3, x_4)$  are clearly measurable from  $(D^2[0, \infty), \mathcal{D}^2)$  to  $(D[0, \infty), \mathcal{D})$ . Thus, since the sum of measurable functions is measurable, it remains to show that both  $b(x_3, x_4)$  and  $d(x_3, x_4)$  are measurable from  $(D^2[0, \infty), \mathcal{D}^2)$  to  $(D[0, \infty), \mathcal{D})$  in order to complete the proof. However, this may be shown in a manner similar to the proof of the measurability of  $\Psi_B^a$  in the proof of Proposition 2 of [11] and we omit the details for the sake of brevity. This completes the proof of the measurability of  $f$ .

We now show that  $f : (\mathbb{R} \times D^3[0, \infty), |\cdot| \times d_{J_1}^3) \mapsto (\mathbb{R} \times D^2[0, \infty), |\cdot| \times d_{J_1}^2)$  is continuous at continuous limit points  $(x_1, x_2, x_3, x_4) \in \mathbb{R} \times D^3[0, \infty)$  such that  $x_2, x_3, x_4 \in C[0, \infty)$ . To begin, it is clear that the functions  $f_1 : (\mathbb{R}, |\cdot|) \mapsto (\mathbb{R}, |\cdot|)$  and  $f_2 : (D[0, \infty), u) \mapsto (D[0, \infty), u)$  are continuous since both of these functions are the identity functions. We now show that the function  $f_3 : (D^2[0, \infty), u^2) \mapsto (D[0, \infty), u)$  is continuous, where we define the maximum metric  $u^2$  for  $(x_1, x_2), (y_1, y_2) \in D^2[0, \infty)$  by

$$u_2((x_1, x_2), (y_1, y_2)) = \max\{u(x_1, y_1), u(x_2, y_2)\}. \quad (8.13)$$

Suppose that  $(x_3^n, x_4^n) \rightarrow (x_3, x_4)$  as  $n \rightarrow \infty$  in  $(D^2[0, \infty), u^2)$  as  $n \rightarrow \infty$ . Then, by (8.13), we have that for each  $T \geq 0$ ,

$$\sup_{0 \leq t \leq T} |x_3^n(t) - x_3(t)| + \sup_{0 \leq t \leq T} |x_4^n(t) - x_4(t)| \rightarrow 0 \text{ as } n \rightarrow \infty.$$

However, this then implies that

$$\begin{aligned}
& \sup_{0 \leq t \leq T} |f_3(x_3^n, x_4^n)(t) - f_3(x_3, x_4)(t)| \\
&= \sup_{0 \leq t \leq T} |x_3^n(t) + \int_0^t x_3^n(t-s)dM(s) + x_4^n(t) + \int_0^t x_4^n(t-s)dM(s) \\
&\quad - (x_3(t) + \int_0^t x_3(t-s)dM(s) + x_4(t) + \int_0^t x_4(t-s)dM(s))| \\
&= \sup_{0 \leq t \leq T} |(x_3^n(t) - x_3(t)) + (x_4^n(t) - x_4(t)) \\
&\quad + \int_0^t (x_3^n(t-s) - x_3(t-s))dM(s) + \int_0^t (x_4^n(t-s) - x_4(t-s))dM(s)| \\
&\leq \sup_{0 \leq t \leq T} |x_3^n(t) - x_3(t)| + \sup_{0 \leq t \leq T} |x_4^n(t) - x_4(t)| \\
&\quad + \sup_{0 \leq t \leq T} \left| \int_0^t (x_3^n(t-s) - x_3(t-s))dM(s) \right| + \sup_{0 \leq t \leq T} \left| \int_0^t (x_4^n(t-s) - x_4(t-s))dM(s) \right| \\
&\leq \sup_{0 \leq t \leq T} |x_3^n(t) - x_3(t)| + \sup_{0 \leq t \leq T} |x_4^n(t) - x_4(t)| \\
&\quad + \sup_{0 \leq t \leq T} \int_0^t |x_3^n(t-s) - x_3(t-s)|dM(s) + \sup_{0 \leq t \leq T} \int_0^t |x_4^n(t-s) - x_4(t-s)|dM(s) \\
&\leq (1 + M(T)) \sup_{0 \leq t \leq T} |x_3^n(t) - x_3(t)| + (1 + M(T)) \sup_{0 \leq t \leq T} |x_4^n(t) - x_4(t)| \\
&\rightarrow 0 \text{ as } n \rightarrow \infty,
\end{aligned}$$

and so the function  $f_3$  is continuous as a map from  $(D^2[0, \infty), u_2)$  to  $(D[0, \infty), u)$ .

Now recall by Chapter 16 of [1] that if  $x_n \rightarrow x$  as  $n \rightarrow \infty$  in  $(D[0, \infty), d_{J_1})$  where  $x \in C[0, \infty)$  is a continuous function, then  $x_n \rightarrow x$  in  $(D[0, \infty), u)$  as well. We therefore have that the function  $f : (\mathbb{R} \times D^3[0, \infty), |\cdot| \times d_{J_1}^3) \mapsto (\mathbb{R} \times D^2[0, \infty), |\cdot| \times d_{J_1}^2)$  defined by (8.11) and (8.12) is continuous at continuous limit points  $(x_1, x_2, x_3, x_4) \in \mathbb{R} \times D^3[0, \infty)$  such that  $x_2, x_3, x_4 \in C[0, \infty)$ . This completes the proof.  $\square$

## References

- [1] P. Billingsley. *Convergence of Probability Measures*. John Wiley and Sons, New York, 1999.
- [2] A. Borovkov. On limit laws for service processes in multi-channel systems. *Siberian Journal of Mathematics*, 8:983–1004, 1967.

- [3] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. John Wiley & Sons, Boca Raton, 1992.
- [4] S. Halfin and W. Whitt. Heavy traffic limits for queues with many exponential servers. *Operations Research*, 29:567–588, 1981.
- [5] J. M. Harrison and R. J. Williams. Brownian models of queueing networks with homogeneous customer populations. *Stochastics*, 22:77–115, 1987.
- [6] D. L. Iglehart. Limit diffusion approximations for the many server queue and the repairman problem. *J. of Applied Probability*, 2:429–441, 1965.
- [7] D. L. Iglehart and W. Whitt. Multiple Channel Queues in Heavy Traffic I. *Advances in Applied Probability*, 2:150–177, 1970.
- [8] S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes*. Academic Press, New York, 1975.
- [9] E. Krichagina and A. Puhalskii. A heavy traffic analysis of a closed queueing system with a GI/ $\infty$  service center. *Queueing Systems*, 25:235–280, 1997.
- [10] A. Puhalskii. On the invariance principle for the first passage time. *Mathematics Operations Research*, 19:946–954, 1994.
- [11] J. E. Reed. The G/GI/N queue in the Halfin-Whitt regime I: Infinite server queue system equations. *Submitted to Annals of Applied Probability*.
- [12] M. I. Reiman. Open queueing networks in heavy traffic. *Mathematics of Operations Research*, 9(3):441–458, 1984.
- [13] S. M. Ross. *Stochastic Processes*. John Wiley & Sons, New York, 1996.
- [14] W. Whitt. *Stochastic-Process Limits*. Springer, New York, 2002.