

# Hazard Rate Scaling for the $GI/M/n + GI$ Queue

Josh Reed  
Stern School of Business  
New York University  
New York, NY

Tolga Tezcan  
University of Illinois  
Department of Industrial and  
Enterprise Systems Engineering  
Urbana, IL

December 10, 2009

## Abstract

We obtain a heavy-traffic limit for the  $GI/M/n + GI$  queue which includes the entire abandonment distribution. Our main approach is to scale the hazard rate function in an appropriate way such that our resulting diffusion approximation contains the entire hazard rate function. We then show through numerical studies that for various key performance measures, our approximations outperform those commonly used in practice which only involved the abandonment distribution through the value of its density at the origin. The robustness of our results is also demonstrated by applying them to solving constraint satisfaction problems arising in the context of telephone call centers.

## 1 Introduction

In this paper, we study multi-server queues with abandonment. Many recent results in the asymptotic analysis of queueing system with abandonment (see for instance [24, 17, 7] and the references therein) seem to suggest that for both single server and multi-server queues with a utilization close to one the density of the abandonment distribution at the origin plays a prominent role in determining overall system performance. Unfortunately, however, approximations based solely on the value of the density of the abandonment distribution at the origin suffer from the potential drawback that the density of a distribution at the origin is a rather un-robust statistic. In order to provide a simple demonstration of this fact consider the following example.

Consider two  $M/M/n + GI$  queues both having an identical arrival rate of 200 customers per unit time and an equal number of servers set equal to 200. Assume that the service rate in both systems is 1 and that the patience time distribution in the first system is exponential with rate 2. In the second system, there are two classes of customers, 90% of whom are very patient, their patience time is exponentially distributed with rate 0.01 and 10% of whom are very impatient, their patience time is exponentially distributed with rate 19.91. Both the patience time distribution in system one and the patience time distribution in system two have the same density at the origin and consequently based off of the approximations provided in [24], these systems will behave identically in the heavy traffic limit considered in [24]. However, direct calculations yield that the average waiting time in steady state in the second system may be up to 70% larger than that in first

system. In addition, the probability of abandonment and probability that a customer waits for service are approximately 11% and 18% less in the second system, respectively.

As the previous example illustrates, an approximation which only incorporates the density of the abandonment distribution at zero may make considerably large approximation errors relative to the actual performance of the system. Moreover, it is well known that in general the hazard rate function of the abandonment distribution plays a pivotal role in characterizing overall system performance [1]. In [22] an asymptotic regime is introduced which appears to overcome many of the above mentioned difficulties and in which the hazard rate function of the abandonment distribution plays a prominent role in the limit. In the present paper, the asymptotic regime considered in [22] is extended to the  $GI/M/n+GI$  queue in the many-server heavy traffic regime in order to obtain new approximations to the steady-state performance of the system. These approximations are shown to be considerable improvements upon those previously obtained. Moreover, we demonstrate the robustness of our results by applying them to solve constraint satisfaction problems in the design of telephone call centers.

Specifically speaking, in the present paper we consider a sequence of  $GI/M/n+GI$  queues in the many-server heavy traffic regime, or more specifically in the QED regime, see [11]. In this regime, both the arrival rate and the number of servers tend to infinity in such a way that the system reaches heavy traffic in the limit. The service time distribution is assumed to be fixed. In the standard QED regime analysis, the abandonment time distribution is also assumed to be fixed, it does not depend on the arrival rate or the system size [24, 17, 7]. On the contrary, in the present paper we also scale the abandonment distribution by setting the hazard rate in the  $n^{th}$  system equal to  $h(\sqrt{n}x)$ , where we recall the number of servers is equal to  $n$ . This allows us to establish a limiting process where the entire abandonment distribution plays a role. Note that for the exponential distribution the hazard rate is constant so that our scaling reduces to the standard QED regime. We first establish diffusion limits for this system and then obtain the stationary distribution of the limiting diffusion process. Using this stationary distribution we propose approximations for certain performance measures. We then test the quality of our suggested approximations via extensive numerical experiments. We show that even when the offered load is as high as 500, simply using the value of the density at zero does not provide satisfactory approximations. In a certain setting, the suggested approximations in the literature may be off by as much as 400%, whereas our approximations in general have relative error of less than 10%. In addition, we use our approximations to find the optimal staffing levels to satisfy a quality of service constraint as in [16]. Our numerical experiments illustrate that solutions based on our approximations are more exact than those provided in [16].

The reasoning behind why just using the value of the abandonment distribution at zero will not always provide a good approximation may be explained as follows. Consider an  $GI/M/n+GI$  queue with an arrival rate  $\lambda$  and service rate  $\mu$  and let  $R = \lambda/\mu$  denote the offered load. It is well known that for large  $R$ , the average waiting time in this system is  $O(1/\sqrt{R})$ . As a consequence, the abandonment behavior of those customers who do abandon is determined by the values of the hazard rate function between zero and  $O(1/\sqrt{R})$ . As  $R$  tends to  $\infty$ , this interval becomes smaller and smaller and hence using an approximation based on the value of the hazard rate function (or the density) of the abandonment time distribution at zero at first glance appears to be quite intuitive. However, if the value of the hazard rate function around zero is rapidly changing, it may take a large  $R$  in order for its value at zero to serve as a good approximation for the hazard rate function of the abandoning customers. Consider for example the hyper-exponential distribution mentioned above. The value of its hazard rate function decreases from 2 to 1 in about 0.04 time

units. The offered load must be over 625 before the order of average waiting times is smaller than 0.04. If we consider a hyper-exponential distribution with rates 1 and 200, the required offered load will be more than 2,500. However, by scaling the hazard rate function we capture the effect of this phenomena and hence even for small  $R$  our method provides reliable approximations.

The rest of the paper is now organized as follows. We close this section with a review of the related literature and summarize the notation and terminology used throughout the rest of the paper. In Section 2, we introduce the mathematical model and our asymptotic framework. We present our main results in Section 3. In Section 4, we explain when our approximations are expected to perform better than those commonly found in the literature. Numerical experiments are presented in Section 5. Sections 6 through 9 are devoted to the proofs of our main results. Specifically, our proof technique relies upon the the Continuous Mapping Theorem and is detailed at the outset of Section 6. In essence, the main idea is to first transform the system equations into a suitable form. We then identify a key process as being a martingale and subsequently apply the martingale convergence theorem as well as the Functional Weak Law of Large Numbers in order to prove our heavy traffic fluid and diffusion limits.

## 1.1 Literature review

The earliest work on the  $M/M/n + M$  queue appears in [19, 20]. Independently, the  $GI/M/n + GI$  queue was analyzed in [1] and [14]. In [18], several approximations to the probability of abandonment in the  $GI/M/n + GI$  queue are developed and tested via simulation. [3, 4] considered the more general  $M(k)/M(k)/n + GI$  system where the arrival and service rates are allowed to depend on the number of calls in the system. A more detailed survey may be found in [24].

Our asymptotic analysis is similar to the QED regime first introduced in [13] for the  $M/M/N$  queue. In [12], the QED regime for the Erlang-A model with exponential abandonment is studied, establishing results that are analogous to [13]. Based on [1], [24] studied the asymptotic behavior of various performance measures when the arrival rate and the number of servers grow to infinity. More specifically, they established formulas for the limiting delay probability, expected waiting time and probability of abandonment among other measures in the quality and efficiency driven (QED) and efficiency driven (ED) regimes. We compare our results to theirs in the QED regime. In addition, [24] studied the staffing problem for constraint for the quality of service in [16]. [5] establishes the limit of the queue length processes for the  $G/Ph/N + GI$  queue in the QED regime. Here  $Ph$  indicates that the service time distribution is of phase-type. [17] establishes process level limits for the  $G/GI/N + GI$  queue.

In [25] and [26], the  $M/M/1 + M$  and  $G/G/1 + GI$  queues, respectively, are analyzed in the conventional heavy traffic regime where customer patience times do not change. Similar to [24], only the density of the abandonment distribution at zero appears in the limit. In order to study the effect of the entire distribution, [22] studied the  $GI/GI/1 + GI$  queue in the conventional heavy traffic regime where the hazard rate function of the abandonment distribution is scaled appropriately. Our approach is similar to theirs.

## 1.2 Notation

All random variables and processes are defined on a common probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  unless otherwise specified. The space of functions  $f : \mathbb{R}_+ \rightarrow \mathbb{R}^d$  that are right-continuous on  $[0, \infty)$  and have left limits in  $(0, \infty)$  is denoted by  $\mathbb{D}(\mathbb{R}_+, \mathbb{R}^d)$  or simply  $\mathbb{D}^d$ ; similarly, with  $T > 0$ , the space of

functions  $f : [0, T] \rightarrow \mathbb{R}^d$  that are right-continuous on  $[0, T)$  and have left limits in  $(0, T]$  is denoted by  $\mathbb{D}([0, T], \mathbb{R}^d)$ . For  $f \in \mathbb{D}$ ,  $f(t-)$  denotes its left limit at  $t > 0$ . For a sequence of random elements  $\{X^n, n \in \mathbb{N}\}$  taking values in a metric space, we write  $X^n \Rightarrow X$  to denote the convergence of  $X^n$  to  $X$  in distribution. Each stochastic process whose sample paths are in  $\mathbb{D}^d$  is considered to be a  $\mathbb{D}^d$ -valued random element. The space  $\mathbb{D}^d$  is assumed to be endowed with the Skorohod  $J_1$ -topology (see [2]). Given  $x \in \mathbb{R}$ , we set  $x^+ = \max\{x, 0\}$  and  $x^- = \max\{-x, 0\}$ .

## 2 The Model

The model which we consider is the  $GI/M/n + GI$  queue with arrival rate  $\lambda$ , service rate  $\mu$  and abandonment distribution  $F$ . We assume that the abandonment distribution is absolutely continuous with hazard rate  $h$ . Moreover, we assume that  $h$  is bounded and we denote its bound by  $\|h\|_\infty$ . Let  $N_A$  be a renewal arrival process with sequence of arrival times

$$\tau_i = \inf\{t \geq 0 : N_A(t) \geq i\},$$

for  $i \geq 1$ . We assume that the sequence of interarrival times of the renewal process have mean one with standard deviation  $\sigma_A$ . Also, let  $N_D$  be an independent, rate one Poisson process. Finally, we denote by  $\{p_i, i \geq 1\}$  the sequence of i.i.d. random variables with CDF  $F$  which represent the patience times of the customers arriving to the system. In particular, the  $i^{th}$  customer arriving to the system will wait no more than  $p_i$  units of time before receiving service. Otherwise, he will abandon.

We now define the total number of customers in the system process  $Q = \{Q(t), t \geq 0\}$  as well as the virtual waiting time process  $W = \{W(t), t \geq 0\}$  for the  $GI/M/n + GI$  queue with arrival rate  $\lambda$  and service rate  $\mu$  to be the solution to the system of equations

$$Q(t) = N_A(\lambda t) - N_D\left(\mu \int_0^t (Q(s) \wedge n) ds\right) - R(t), \quad (1)$$

$$R(t) = \sum_{i=1}^{N_A(\lambda t)} 1\{p_i \leq W(\tau_i-)\} 1\{p_i \leq t - \tau_i\}, \quad (2)$$

and

$$\begin{aligned} & W(t) \quad (3) \\ = & \inf \left\{ u \geq 0 : \left( N_D\left(\mu \int_0^{t+u} (Q(s) \wedge n) ds\right) - N_D\left(\mu \int_0^t (Q(s) \wedge n) ds\right) \right) \right. \\ & \left. + \left( \sum_{i=1}^{N_A(\lambda t)} 1\{p_i \leq W(\tau_i-)\} 1\{p_i \leq (t+u) - \tau_i\} - \sum_{i=1}^{N_A(\lambda t)} 1\{p_i \leq W(\tau_i-)\} 1\{p_i \leq t - \tau_i\} \right) \right\} \\ \geq & (Q(t) - n)^+. \end{aligned}$$

It can be shown that this system of equations possesses a unique solution. Moreover, the process  $R = \{R(t), t \geq 0\}$  represents the cumulative number of customers who have abandoned from the system at each point in time.

## 2.1 Asymptotic framework

We analyze the  $GI/M/n + GI$  queue in the many-server heavy-traffic regime explained in the following. Consider a sequence of  $GI/M/n + GI$  queues indexed by the number of servers  $n$ . In the  $n^{\text{th}}$  system we assume that the arrival rate is  $\lambda^n$ , i.e. the arrival process is  $N_A(\lambda^n e)$ , and that the service rate is  $\mu$ . We could easily allow the service rate to vary with  $n$  as well, however for the sake of brevity in our proofs we do not. Finally, assume that the hazard rate of the abandonment distribution for the customers in the  $n^{\text{th}}$  system,  $h^n$ , is defined by setting  $h^n(x) = h(\sqrt{n}x)$  for  $x \geq 0$ , as was done in [22]. Our convention is to superscript all quantities associated with the  $n^{\text{th}}$  system by the letter  $n$ .

Let  $Q_0^n$  denote the initial number of customer in the  $n$ th system. We define

$$\tilde{Q}_0^n = \frac{Q_0^n - n}{\sqrt{n}}, \quad (4)$$

$$\tilde{Q}^n(t) = \frac{Q^n(t) - n}{\sqrt{n}}, \quad (5)$$

$$(6)$$

We assume that

$$n^{-1}\lambda^n \rightarrow \lambda, \quad \text{as } n \rightarrow \infty, \quad (7)$$

where  $\lambda = \mu$  and

$$n^{1/2}(n^{-1}\lambda^n - \mu) \rightarrow -\beta\mu \quad \text{as } n \rightarrow \infty. \quad (8)$$

We also assume that

$$\tilde{Q}_0^n \Rightarrow \tilde{Q}_0, \quad \text{as } n \rightarrow \infty, \quad (9)$$

for some random variable  $\tilde{Q}_0$ .

## 3 Main results

The following is the main result of the paper. It provides a diffusion limit for the  $GI/M/n + GI$  queue in the Halfin-Whitt regime under the hazard rate scaling.

**Theorem 3.1.** *Assume that (7)–(9) hold. Then  $\tilde{Q}^n \Rightarrow \tilde{Q}$  as  $n \rightarrow \infty$ , where  $\tilde{Q}$  is the unique, strong solution to the stochastic differential equation*

$$\tilde{Q}(t) = \sqrt{\mu(1 + \sigma_A^2)}\tilde{B}(t) - \mu\beta t - \mu \int_0^t \tilde{Q}^-(s)ds - \mu \int_0^t \left( \int_0^{\tilde{Q}^+(s)} h(u)du \right) ds,$$

for  $t \geq 0$ .

We now provide the steady-state distribution the limiting process  $\tilde{Q}$  above.

**Proposition 3.2.** *If  $\int_0^\infty h(u) > -\beta$ , then  $\tilde{Q}$  possess a stationary distribution  $\pi$  whose density given by*

$$\pi(dx) = C \exp\left(\frac{-2}{1+\sigma_A^2} \left(\beta x + \int_0^x \left(\int_0^u h(v)dv\right) du\right)\right) dx, \text{ for } x \geq 0$$

and

$$\pi(dx) = C \exp\left(\frac{-2}{1+\sigma_A^2} (\beta x + x^2/2)\right) dx, \text{ for } x \leq 0,$$

where

$$C = \left(\int_0^\infty \left(\exp\left(\frac{-2}{1+\sigma_A^2} \left(\beta x + \int_0^x \left(\int_0^u h(v)dv\right) du\right)\right) + \exp\left(\frac{-2}{1+\sigma_A^2} (\beta x + x^2/2)\right)\right) dx\right)^{-1}.$$

### 3.1 Approximations

Using the stationary distribution of the limiting process  $\tilde{Q}$ , we may now obtain the following approximations. Given the hazard rate function  $h$  for the abandonment distribution, let  $h^n = h(x/\sqrt{n})$ . Given the arrival rate  $\lambda^n$  and the number of servers  $n$ , let the pdf  $\pi^n$  be defined as follows;

$$\pi^n(dx) = C \exp\left(\frac{-2}{1+\sigma_A^2} \left(\beta^n x + \int_0^x \left(\int_0^u h^n(v)dv\right) du\right)\right) dx, \text{ for } x \geq 0$$

and

$$\pi^n(dx) = C \exp\left(\frac{-2}{1+\sigma_A^2} (\beta^n x + x^2/2)\right) dx, \text{ for } x \leq 0,$$

where

$$C = \left(\int_0^\infty \left(\exp\left(\frac{-2}{1+\sigma_A^2} \left(\beta^n x + \int_0^x \left(\int_0^u h^n(v)dv\right) du\right)\right) + \exp\left(\frac{-2}{1+\sigma_A^2} (\beta^n x + x^2/2)\right)\right) dx\right)^{-1},$$

and

$$\beta^n = -\mu^{-1}(\lambda^n - n\mu)/\sqrt{n}.$$

We now approximate the stationary distribution of the queue length in the  $n$ th system by  $\sqrt{n}\pi^n$ . We can also approximate the steady state distribution of the waiting time by  $\pi^n/(\sqrt{n}\mu)$ , see Proposition 8.3 below.

We now provide approximations for certain key performance measures. The performance measures we focus on is the probability that all the servers are busy, which we denote by  $P(W > 0)$ , the expected waiting time, which we denote by  $EW$ , and the steady-state probability of abandonment, which we denote by  $PAb$ . If arrivals are Poisson, then by the PASTA property,  $P(W > 0)$  is also equal to the probability that an arriving customer will have to wait in steady state. Our asymptotic results suggest the following approximations;

$$\begin{aligned}
& P\{W > 0\} \\
& \approx \frac{\int_0^\infty \exp\left(\frac{-2}{1+\sigma_A^2} (\beta^n x + \int_0^x (\int_0^u h^n(v)dv)) du\right) dx}{\int_0^\infty \exp\left(\frac{-2}{1+\sigma_A^2} (\beta^n x + \int_0^x (\int_0^u h^n(v)dv)) du\right) dx + \int_{-\infty}^0 \exp\left(\frac{-2}{1+\sigma_A^2} (\beta^n x + x^2/2)\right) dx} \quad (10)
\end{aligned}$$

$$EQ \approx \sqrt{n}C \int_0^\infty x \exp\left(\frac{-2}{1+\sigma_A^2} (\beta^n x + \int_0^x (\int_0^u h^n(v)dv)) du\right) dx \quad (11)$$

$$EW \approx \sqrt{n}EQ/\mu \quad (12)$$

$$P(Ab) \approx 1 - \mu(N - EB)/\lambda, \quad (13)$$

where

$$EB = n + \sqrt{n}C \int_{-\infty}^0 x \exp\left(\frac{-2}{1+\sigma_A^2} (\beta^n x + x^2/2)\right) dx \quad (14)$$

is the expected number of busy agents.

## 4 Intuition

We now provide a bit more of an intuitive explanation for when one should expect the approximations provided in this paper to outperform than those given in [24]. In [24], the approximations to the queue length process is provided by a diffusion process with a linear drift coefficient given by  $b(x) = \mu\beta + f(0)x$  where  $\beta$  is a constant which may be determined as a function of the capacity imbalance of the system and  $f(0)$  is the density of the abandonment distribution at zero.

In the present paper, our drift function is given by  $b(x) = \mu\beta + \int_0^x h(u)du$ . However, note that so long as the abandonment distribution does not have an atom at zero, we have that  $h(0) = f(0)$ . Thus, assuming that the hazard rate of the abandonment distribution is differentiable at least  $k \geq 1$  times, we obtain by Taylor's Theorem [8] expanding about the point zero that

$$h(x) = f(0) + h'(0)x + \sum_{j=2}^k h^{(j)}(0) \frac{x^j}{j!} + R_k(x),$$

where  $R_k(x)$  is a remainder term. Substituting into our drift function, we then see that our drift function may be written as

$$\begin{aligned}
b(x) &= \mu\beta + \int_0^x \left( f(0) + h'(0)u + \sum_{j=2}^k h^{(j)}(0) \frac{u^j}{j!} + R_k(u) \right) du \\
&= \mu\beta + f(0)x + \frac{1}{2}h'(0)x^2 + \sum_{j=2}^k h^{(j)}(0) \frac{x^{j+1}}{(j+1)!} + \int_0^x R_k(u)du.
\end{aligned}$$

Comparing with the original drift function  $b(x) = \mu\beta + f(0)x$ , one then sees that one should expect our approximations to perform particularly well in the case when the magnitude of the derivative (and higher orders thereof) of the hazard rate function at zero is large. Indeed, we find out that is the case in the Section that follows.

## 5 Numerical Experiments

In this Section we test the quality of our proposed approximations via extensive numerical experiments. We first focus on the performance measures  $P(W > 0)$ ,  $EW$ , and  $PAb$ . Then, in Section 5.3, we use our approximations to find the minimum staffing levels to satisfy a service quality constraint. We compare our approximations with those in [24] and [16]. We refer to the approximations appearing in [24] as the *Z&M* approximations.

### 5.1 Discussion of the experiments

We consider four different different  $M/M/N + G$  queues with varying patience time distributions and with the number of servers in the systems ranging from 10 to 500. We also consider systems with different levels of load. Specifically, we consider  $\beta = -1, 0$  and  $1$ . The varying patience time distributions are chosen to illustrate some of the points alluded to in the introduction. In two of our experiments, we use patience times with a hyper-exponential distribution. A random variable  $X$  is said to have a hyper-exponential distribution with rates  $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_K)$  and probabilities  $q = (q_1, \dots, q_K)$ , with  $\sum_k q_k = 1$  and  $q_k, \gamma_k \geq 0$ , for  $k = 1, \dots, K$ , if its pdf  $f_X$  is given by

$$f_X(x) = \sum_{i=k}^K f_{Y_k}(x)q_k, \quad (15)$$

where  $Y_i$  is an exponentially distributed random variable with rate parameter  $\gamma_k$  and  $f_{Y_k}$  is its pdf.

The first distribution we consider is a hyper-exponential distribution with rates  $(1, 2)$  with probabilities  $(0.5, 0.5)$ . We choose this distribution as a base case to compare our approximations with the *Z&M* approximations in a setting where the hazard rate of the abandonment time distribution around zero is not changing fast. The second and the third distributions are chosen to illustrate the advantage of our approximation when this is not the case. Specifically, the second distribution is also a hyper-exponential distribution with rates  $(1, 200)$  with probabilities  $(0.9, 0.1)$ . The derivative of the hazard rate of this distribution at zero is negative. The third patience time distribution is a distribution with the following hazard rate function;

$$h(x) = \begin{cases} h_0 + \frac{K-h_0}{b}x & \text{if } 0 \leq x \leq b \\ K & \text{if } x > b, \end{cases} \quad (16)$$

in which the derivative at zero is clearly positive. In order to make the comparisons similar to experiment 1 we set  $h_0 = 1.5$ . We also set  $b = 0.1$  and  $K = 100$ . In Figure 1, we plot the hazard rate functions of these three distributions in addition to the associated cdf of the increasing hazard rate function. Note that, because the *Z&M* approximations only depend on the value of the hazard rate function at zero, the approximations of *Z&M* for the first and the third experiments are the same.

In our final experiment, we consider an Erlang distribution with shape  $k = 2$  and rate  $\gamma = 4$ . The hazard rate function of an Erlang distribution vanishes at zero and hence the *Z&M* approximations are not directly applicable. In [24], an alternative method is suggested when this is the case that involves using the value of the smallest derivative of the hazard rate function that does not vanish at zero. For our approach no such modification is needed since we use the entire hazard rate function  $h$ .

In all of our experiments the service and interarrival times have exponential distributions and the service rate equals 1. The results of these experiments are given in Tables 1 through 9. For each different patience time distribution, we have three tables for three different staffing levels  $\beta = -1, 0$ , and 1.

## 5.2 Discussion of the results for performance measures

We begin by noting that when the patience time has a hyper-exponential distribution with rates (1, 2) and probabilities (0.5, 0.5), both approximations seem to provide near exact values even for relatively small systems, see Tables 1–3. This is expected since the hazard rate function is not changing rapidly around zero.

The next patience time distribution we consider is a hyper-exponential distribution with rates (1, 200) and probabilities (0.9, 0.1). We again consider the average absolute error from the exact values for  $\beta = 1, 0, -1$ , excluding the case of  $N = 10$ . We note that our approximation continues to provide fairly accurate values, see Tables 4–5. In terms of  $P(W > 0)$ , the Z&M approximations are off by 44% while our approximations are only off by 3.3% on the average. In terms of  $EW$ , the Z&M approximations are off by 78%, our approximation is off by 4.7%. In approximating the probability of abandonment the Z&M approximation performs better than previously, with a relative error of 21%, however our approximations maintain a relative error of only 4.4%. Clearly in this case our approximations appear to be performing better than those of Z&M.

The third patience time distribution we consider has the hazard rate function given by (16). Note that when  $h_0 = 1.5$ , the Z&M approximations are the same as for the first patience time distribution (hyper-exponential with rate (1, 2) and probabilities (0.5, 0.5)) since the value of the density at zero is the same in both cases. A quick review of Tables 7–9 reveals that the Z&M approximations do not appear to be very accurate in this case, even when the system size is large. To illustrate this, we again compute average absolute relative errors, excluding  $N = 10$ , for three different staffing levels for all three performance measures. In terms of  $P(W > 0)$ , the Z&M approximation is off by 80%, for  $EW$ , it is off by 463% and in terms of,  $PAb$ , it is off by 28%. On the other hand, for our approximations, these errors are 6.9%, 2.8% and 3.3%, respectively. Again, this experiment shows that our approximations perform significantly better than the Z&M approximations when the hazard rate function changes fast around zero.

Finally, the improvements of our approach is also evident when we consider patience times with an Erlang distribution, see Tables 13–15. For example, in terms of  $EW$ , the average absolute error of our approximation compared to the exact results, excluding the cases for  $N = 10$ , when the system load is  $\beta = -1, 0, 1$  is around 2.1%. For the same performance measure the Z&M approximation is off by 17% on average.

## 5.3 Optimal staffing levels for constraint satisfaction

In this Subsection, we run some numerical experiments in order to find the optimal staffing levels to satisfy a constraint on service quality. The objective is to find the minimum staffing level  $N$  that guarantees a certain service level for a fixed arrival rate. The quality of service measure we focus on is the probability of delay,  $P(W > 0)$ .

In order to find the optimal staffing level, we perform a simple search. First we fix  $\lambda$  and the service level requirement  $\alpha$ . We then start from a small  $N$  and find  $P(W > 0)$  for the fixed level of  $\lambda$  and staffing level  $N$ . If  $P(W > 0)$  is not below the desired level  $\alpha$ , we increase  $N$  by 1 and repeat

N	$PW > 0$			$EW$			$PAb$		
	Exact	Diffusion	HRS	Exact	Diffusion	HRS	Exact	Diffusion	HRS
10	0.4996	0.4495	0.4524	5.6201	5.5560	5.6817	0.1367	0.1389	0.1382
50	0.4721	0.4495	0.4508	2.5118	2.4847	2.5094	0.062	0.0621	0.062
100	0.4651	0.4495	0.4504	1.7674	1.7570	1.7693	0.0438	0.0439	0.0439
200	0.4612	0.4495	0.4501	1.2498	1.2424	1.2485	0.0311	0.0311	0.031
500	0.4565	0.4495	0.4499	0.788	0.7857	0.7882	0.0196	0.0196	0.0196

Table 1: Hyper-exponential with  $p = (0.5, 0.5)$  and  $\gamma = (1, 2)$ ,  $\beta = 0$

N	$PW > 0$			$EW$			$PAb$		
	Exact	Diffusion	HRS	Exact	Diffusion	HRS	Exact	Diffusion	HRS
10	0.1452	0.1404	0.1460	1.2661	1.2698	1.2863	0.0310	0.0317	0.0461
50	0.1465	0.1404	0.1457	0.5733	0.5679	0.5711	0.0142	0.0142	0.0165
100	0.1464	0.1404	0.1456	0.4049	0.4015	0.4032	0.0101	0.01	0.0111
200	0.1462	0.1374	0.1455	0.2858	0.2839	0.2847	0.0071	0.0066	0.0076
500	0.1459	0.1404	0.1455	0.1804	0.1722	0.1799	0.0045	0.0043	0.0047

Table 2: Hyper-exponential with  $p = (0.5, 0.5)$  and  $\gamma = (1, 2)$ ,  $\beta = -1$

N	$PW > 0$			$EW$			$PAb$		
	Exact	Diffusion	HRS	Exact	Diffusion	HRS	Exact	Diffusion	HRS
10	0.7952	0.7755	0.7806	11.3858	14.1404	14.6407	0.2755	0.3535	0.2679
50	0.7839	0.7755	0.7777	5.6986	6.3238	6.4212	0.1403	0.1581	0.1384
100	0.7814	0.7755	0.7771	4.1484	4.4716	4.5199	0.1026	0.1118	0.1016
200	0.7796	0.7755	0.7766	2.9963	3.1619	3.1859	0.0743	0.0790	0.0738
500	0.7781	0.7755	0.7762	1.9320	1.9998	2.0093	0.0481	0.0500	0.0478

Table 3: Hyper-exponential with  $p = (0.5, 0.5)$  and  $\gamma = (1, 2)$ ,  $\beta = 1$

N	$PW > 0$			$EW$			$PAb$		
	Exact	Diffusion	HRS	Exact	Diffusion	HRS	Exact	Diffusion	HRS
10	0.4886	0.1795	0.4399	5.5084	0.298	5.7445	0.1397	0.1038	0.1413
50	0.4086	0.1795	0.382	1.8414	0.1864	1.8818	0.0696	0.0649	0.0697
100	0.3679	0.1795	0.3485	1.0599	0.1448	1.0802	0.0518	0.0505	0.052
200	0.3282	0.1795	0.3126	0.5841	0.11	0.5904	0.0387	0.0383	0.0388
500	0.2779	0.1795	0.2676	0.2513	0.0744	0.2526	0.0261	0.0259	0.0261

Table 4: Hyper-exponential with  $p = (0.9, 0.1)$  and  $\gamma = (1, 200)$ ,  $\beta = 0$

N	$PW > 0$			$EW$			$PAb$		
	Exact	Diffusion	HRS	Exact	Diffusion	HRS	Exact	Diffusion	HRS
10	0.1407	0.0626	0.1385	1.1893	0.1879	1.2219	0.0335	0.0655	0.0505
50	0.1255	0.0626	0.1207	0.4157	0.0840	0.4148	0.0186	0.0293	0.0218
100	0.1154	0.0626	0.1109	0.2470	0.0594	0.2455	0.0145	0.0207	0.0161
200	0.1045	0.0626	0.1006	0.1414	0.0420	0.1402	0.01045	0.0146	0.0120
500	0.0909	0.0626	0.0881	0.0654	0.0266	0.06486	0.0078	0.0093	0.0082

Table 5: Hyper-exponential with  $p = (0.9, 0.1)$  and  $\gamma = (1, 200)$ ,  $\beta = -1$

N	$PW > 0$			$EW$			$PAb$		
	Exact	Diffusion	HRS	Exact	Diffusion	HRS	Exact	Diffusion	HRS
10	0.7909	0.3344	0.7821	11.8884	1.2251	16.2358	0.2763	0.4268	0.2677
50	0.7158	0.3344	0.7090	4.5527	0.5479	5.4050	0.1455	0.1909	0.1428
100	0.6663	0.3344	0.6587	2.7067	0.3874	3.0844	0.1087	0.1350	0.1072
200	0.6063	0.3344	0.5979	1.4942	0.2740	1.6453	0.0808	0.0954	0.0800
500	0.5213	0.3344	0.5127	0.6201	0.1733	0.6577	0.0541	0.0604	0.0538

Table 6: Hyper-exponential with  $p = (0.9, 0.1)$  and  $\gamma = (1, 200)$ ,  $\beta = 1$

N	$PW > 0$			$EW$			$PAb$		
	Exact	Diffusion	HRS	Exact	Diffusion	HRS	Exact	Diffusion	HRS
10	0.2716	0.4495	0.1578	0.5163	5.5560	0.4001	0.199	0.1389	0.2125
50	0.2305	0.4495	0.1945	0.3069	2.4847	0.285	0.0901	0.0621	0.0909
100	0.2344	0.4495	0.2119	0.2548	1.7570	0.2447	0.0627	0.0439	0.0629
200	0.2446	0.4495	0.2299	0.2138	1.2424	0.2091	0.0434	0.0311	0.0434
500	0.2633	0.4495	0.2547	0.1701	0.7857	0.1682	0.0266	0.0196	0.0266

Table 7: Increasing hazard rate,  $\beta = 0$

N	$PW > 0$			$EW$			$PAb$		
	Exact	Diffusion	HRS	Exact	Diffusion	HRS	Exact	Diffusion	HRS
10	0.0879	0.1454	0.0560	1.3874	1.2698	0.1334	0.0634	0.7311	0.0997
50	0.0779	0.1454	0.0687	0.0992	0.5679	0.0928	0.0285	0.0142	0.0328
100	0.0799	0.1454	0.0746	0.0816	0.4015	0.0787	0.0194	0.01	0.0213
200	0.0839	0.1454	0.0807	0.0677	0.2839	0.0663	0.0132	0.0071	0.0140
500	0.0906	0.1454	0.0889	0.0528	0.1796	0.0523	0.0078	0.004	0.0081

Table 8: Increasing hazard rate,  $\beta = -1$

N	$PW > 0$			$EW$			$PAb$		
	Exact	Diffusion	HRS	Exact	Diffusion	HRS	Exact	Diffusion	HRS
10	0.4454	0.7755	0.2913	0.8557	14.1404	0.7860	0.3312	0.3535	0.3297
50	0.4041	0.7755	0.3559	0.5605	6.3238	0.5650	0.1679	0.1581	0.1658
100	0.4152	0.7755	0.3859	0.4786	1.7570	0.4877	0.1215	0.1118	0.1202
200	0.4351	0.7755	0.4169	0.4107	1.2424	0.4191	0.0870	0.0790	0.0863
500	0.4688	0.7755	0.4588	0.3347	1.99987	0.3404	0.0553	0.0500	0.0550

Table 9: Increasing hazard rate,  $\beta = 1$

$\lambda$	$\alpha$		
	0.1	0.5	0.9
10	(15,14,14)	(11,10,10)	(6,5,6)
50	(60,59,59)	(50,50,50)	(40,39,39)
100	(113,113,113)	(100,99,99)	(85,84,84)
200	(219,218,218)	(199,199,199)	(179,178,178)
500	(529,528,528)	(498,497,497)	(465,464,465)
1000	(1041,1040,1040)	(997,996,996)	(951,949,950)

Table 10:  $p = (0.5, 0.5)$  and  $\lambda = (1, 2)$

$\lambda$	$\alpha$		
	0.1	0.5	0.9
10	(15,12,14)	(10,4,10)	(7,1,6)
50	(59,55,58)	(49,36,48)	(39,5,38)
100	(112,107,111)	(96,80,96)	(82,36,81)
200	(215,209,215)	(192,171,191)	(169,109,168)
500	(522,514,521)	(481,454,480)	(438,356,437)
1000	(1028,1020,1027)	(965,935,963)	( 895,797,892)

Table 11:  $p = (0.9, 0.1)$  and  $\lambda = (1, 200)$

$\lambda$	$\alpha$		
	0.1	0.5	0.9
10	(14,14,12)	(7,10,1)	(2,5,1)
50	(57,59,56)	(40,50,36 )	(17,39,1)
100	(109,113,108 )	(86,99,83)	( 51,84,28)
200	(213,218,212)	(182,199,179 )	(133,178,119)
500	(522,528,521)	(475,497,473)	(403,464,394)
1000	(1032,1040,1031)	(969,996,967)	( 874,949,867)

Table 12: Increasing Hazard Rate

N	$PW > 0$			$EW$			$PAb$		
	Exact	Diffusion	HRS	Exact	Diffusion	HRS	Exact	Diffusion	HRS
10	0.4454	0.6675	0.7457	28.8347	24.7671	28.2191	0.0643	0.2070	0.0642
50	0.8205	0.8056	0.8130	20.3402	22.8569	20.1901	0.0211	0.0177	0.0211
100	0.8433	0.8457	0.8387	17.6314	21.3772	17.5513	0.0129	0.0104	0.0129
200	0.8648	0.8775	0.8619	15.31	19.7619	15.2755	0.0078	0.0061	0.0078
500	0.8903	0.9097	0.8888	12.7424	17.5863	12.7244	0.0040	0.0029	0.0040

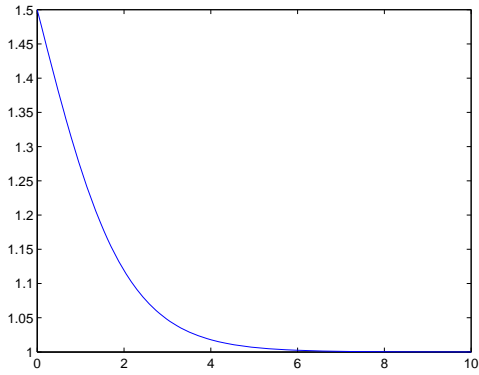
Table 13: Erlang( $k = 2, \gamma = 4$ ),  $\beta = 0$

N	$PW > 0$			$EW$			$PAb$		
	Exact	Diffusion	HRS	Exact	Diffusion	HRS	Exact	Diffusion	HRS
10	0.2157	0.2234	0.7457	3.2921	4.2380	3.6012	0.0030	0.0113	0.0046
50	0.2124	0.2234	0.2212	1.7311	1.8953	1.7964	3.8403e-004	2.0216e-004	4.5660e-004
100	0.2162	0.2234	0.2223	1.2684	1.3402	1.3019	1.3774e-004	3.5738e-005	1.5569e-004
200	0.2187	0.2234	0.2229	0.9171	0.9476	0.9341	4.5573e-005	6.3176e-006	4.9729e-005
500	0.2206	0.2234	0.2232	0.5895	0.5993	0.5963	9.4624e-006	6.3930e-007	1.0005e-005

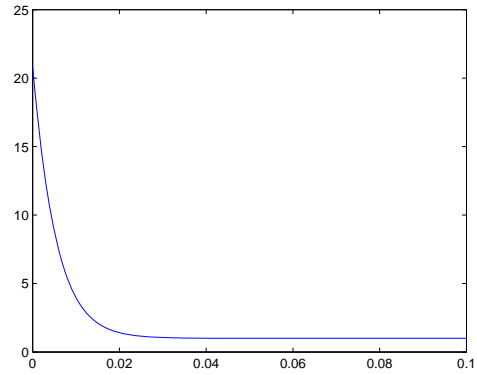
Table 14: Erlang( $k = 2, \gamma = 4$ ),  $\beta = -1$

N	$PW > 0$			$EW$			$PAb$		
	Exact	Diffusion	HRS	Exact	Diffusion	HRS	Exact	Diffusion	HRS
10	0.9902	1.00	0.9907	64.8818	71.3096	73.6288	0.2419	0.3162	0.2414
50	0.9989	1.00	0.9990	52.4479	60.7109	55.5308	0.1240	0.1414	0.1240
100	0.9997	1.00	0.9998	47.5764	56.6453	49.5195	0.0909	0.0100	0.0909
200	1.00	1.00	1.00	43.0605	52.8519	44.2818	0.0660	0.0707	0.0660
500	1.00	0.9097	1.00	37.6727	48.2244	38.3332	0.0428	0.0447	0.0428

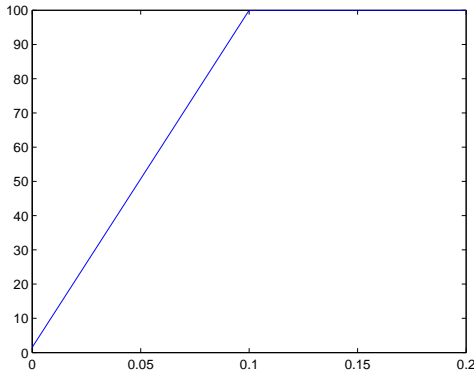
Table 15: Erlang( $k = 2, \gamma = 4$ ),  $\beta = 1$



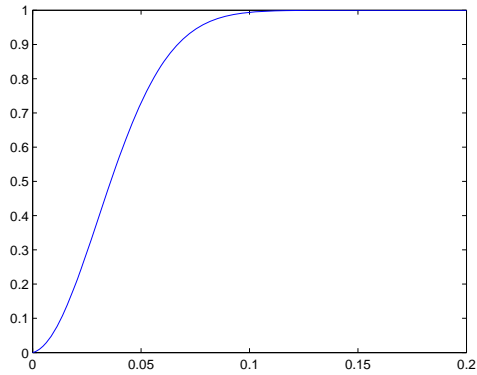
(a) Hazard Rate Function: Hyper-exponential with  $p = (0.5, 0.5)$  and  $\gamma = (1, 2)$



(b) Hazard Rate function: Hyper-exponential with  $p = (0.9, 0.1)$  and  $\lambda = (1, 200)$



(c) Hazard rate function for (16)



(d) CDF for (16)

Figure 1: Hazard rate functions for Experiments 3 and 4

the process until we find the smallest  $N$  that satisfies the service level requirement. We find the exact optimal solution by evaluating  $P(W > 0)$  using the exact values provided in [24]. They also show that  $P(W > 0)$  is decreasing in  $N$ , see [16], and hence our simple procedure finds the optimal solution. In order to compare the performance of the Z&M approximations and ours, we repeat the same procedure to find the minimum  $N$  by evaluating  $P(W > 0)$  using their approximation and then ours. In [16], it is also shown that their approximation for  $P(W > 0)$  is decreasing in  $N$ , validating our optimization procedure. We prove below it is also monotone when we use our approximation.

In our numerical experiments, we consider arrival rates (10, 50, 100, 200, 500, 1000) and service level constraints  $\alpha = (0.1, 0.5, 0.9)$ . We run experiments under the three of the patience time distributions used in the previous section; in the first experiment the patience time distribution is hyper exponential with probabilities  $p = (0.5, 0.5)$  and rates (1, 2), in the second experiment it is hyper exponential with parameters  $p = (0.9, 0.1)$  with rates (1, 200), and in the third experiment we focus on the increasing hazard rate function given in (16). The results for each experiment are

presented in Tables 10–12. In each entry we present results in a vector format  $(i, j, k)$ . The first value is the result of the exact analysis, the second is that obtained using the Z&M approximation and the third is from our approximation.

In order to validate our search procedure, we need to show that the approximations provided in (10)–(13) are monotone. We first start with the case when  $n$  is fixed. This result follows from directly comparing the sample paths of the diffusion limits of two systems with  $\beta_1 < \beta_2$ . and hence its proof has been omitted.

**Lemma 5.1.** *For fixed  $n$ ,  $P\{W > 0\}$ ,  $EQ$ ,  $EW$  and  $P(Ab)$  are increasing in  $\lambda$ .*

As described above, in order to find the optimal staffing levels, we fix first  $\lambda$  and then search for the minimum  $N$  that satisfies the quality of service constraint using our approximation given in (10). However, our main result is based on an asymptotic scaling on  $n$ . This can be modified easily as is prescribed next. In order to obtain an approximation, we consider a sequence of systems indexed by the arrival rate  $\lambda$  where the hazard rate of the patience times is scaled by setting  $h^\lambda(x) = h(\sqrt{\lambda}x)$ . Compared to a scaling based on the number of servers, the new scaling is just a change of the scaling factor and yields similar results in the QED regime, assuming (8) still holds. Now let

$$\check{Q}^\lambda(t) = \frac{Q^\lambda(t) - N^\lambda}{\sqrt{\lambda}}.$$

Assuming that  $\mu = 1$ , we note that  $\check{Q}^\lambda$  has the same weak limit as  $\tilde{Q}^n$ , hence the approximations given in (10)–(13) are also valid for this scaling. We use these approximations and the monotonicity result Lemma 5.1 in order to find the optimal staffing levels.

#### 5.4 Discussion of the results of the staffing experiments

In the light of the numerical experiments in Section 5.1, we expect that when the patience times have a hyper-exponential distribution with rates  $(1, 2)$  and probabilities  $p = (0.5, 0.5)$ , both approximations should perform well in finding optimal staffing levels. This is indeed the case. The average absolute relative error of both methods are around 0.5%, with our approximation providing slightly better results. When we consider patience times with a hyper-exponential distribution with rates  $(1, 200)$  and probabilities  $p = (0.9, 0.1)$ , our approximations perform significantly better in determining the optimal capacity levels. The average absolute relative error with the Z&M method is around 20%, whereas with our method this error is around 0.7%. This again shows the robustness of our approximations to changes in the hazard rate function of the abandonment times.

When the patience times have a distribution with the increasing hazard rate function given in (16), both approximations do not perform well especially when the service quality constraint is  $\alpha = 0.9$  and arrival rate is low, less than 500. This is mainly due to the fact that in these cases  $\beta$  seems to be very high. For example, when  $\lambda = 100$ , the exact solution is  $N = 51$ , giving  $\beta = 6.8$ . This system is hardly in the QED regime. It may be more appropriate to analyze them in the ED regime, see [24]. However, when  $\lambda = 500$ , the exact solution is  $N = 403$ , giving  $\beta = 4.83$ . In this case the staffing solution based on our solution is only off by only 2.2% compared to 15% for the Z&M approximation. Excluding the experiments with  $\alpha = 0.9$ , the average relative absolute error of the staffing solutions found by our approximations is less than 0.3% compared to 7% for the solutions based on the Z&M approximation.

## 6 Preliminary Proofs

The remainder of the paper is now dedicated to the proofs of our main results. In the following Subsection, we provide some algebraic manipulations of the original system equation for the  $GI/M/n + GI$  queue appearing in Section 2. Next, in Subsection 6.2 we rigorously verify that a certain process appearing in the systems equations in Subsection 6.1 is a martingale. Moreover, we identify its quadratic variation. In Section 7, we provide a fluid limit for the queue length process. Here the basic approach is to use a continuous map in conjunction with the Functional Weak Law of Large Numbers for renewal processes and the martingale invariance principle. In Section 8, we provide a version of the Reiman's snapshot principle [23] proving the asymptotic equivalence between the diffusion scaled queue length process and the diffusion scaled virtual waiting time process. Finally, in Section 9, the main result of the paper is proven, Theorem 3.1, which provides a process level limit for the diffusion scaled queue length process of the  $GI/M/n + GI$  queue in the Halfin-Whitt regime, assuming hazard rate scaling of the abandonment distribution.

### 6.1 System equations

We now perform some manipulations on equation (1) for the queue length process. Let  $\hat{N}_A(t) = (N(\lambda t) - \lambda t)$  and set

$$\hat{N}_D(t) = N_D \left( \mu \int_0^t (Q(s) \wedge n) ds \right) - \mu \int_0^t (Q(s) \wedge n) ds.$$

It then follows from (1) that

$$Q(t) = \hat{N}_A(\lambda t) - \hat{N}_D(t) + \lambda t - \mu \int_0^t (Q(s) \wedge n) ds - R(t).$$

Moreover, upon noting that  $n - (Q(s) \wedge n) = (n - Q(s))^+ = -(Q(s) - n)^-$ , we obtain that

$$Q(t) = \hat{N}_A(\lambda t) - \hat{N}_D(t) + (\lambda - n\mu)t - \mu \int_0^t (Q(s) - n)^- ds - R(t). \quad (17)$$

Finally, define

$$\hat{R}(t) = R(t) - \sum_{i=1}^{N_A(\lambda t)} \int_0^{(t-\tau_i) \wedge W(\tau_i^-) \wedge p_i} h(u) du.$$

It then follows from (17) that

$$\begin{aligned} Q(t) &= \hat{N}_A(\lambda t) - \hat{N}_D(t) + \hat{R}(t) + (\lambda - n\mu)t - \mu \int_0^t (Q(s) - n)^- ds \\ &\quad - \sum_{i=1}^{N_A(\lambda t)} \int_0^{(t-\tau_i/\lambda) \wedge W(\tau_i/\lambda^-) \wedge p_i} h(u) du. \end{aligned} \quad (18)$$

This will be our final system equation.

## 6.2 Martingale results

We now provide a martingale result which will be useful for the remainder of the paper. In particular, it is helpful in proving of our fluid limit result in Section 7 as well as our tightness and asymptotic equivalence results in Section 8. Note also that this result is a direct application of Proposition 5.5 of [15]. For each  $t \geq 0$ , let  $\mathcal{F}_t$  be the  $\sigma$ -algebra defined by setting

$$\begin{aligned} \mathcal{F}_t &= \sigma\{N_A(\lambda s), 0 \leq s \leq t\} \vee \sigma\left\{N_D\left(\mu \int_0^s (Q(s) \wedge n) ds\right), 0 \leq s \leq t\right\} \\ &\quad \vee \{1\{p_i \leq W(\tau_i-)\}\} 1\{p_i \leq t - \tau_i\}, 1 \leq i \leq N_A(\lambda t)\}, \end{aligned}$$

and define the filtration  $\mathcal{F} = \{\mathcal{F}_t, t \geq 0\}$ .

**Proposition 6.1.** *The process  $\hat{R} = \{\hat{R}(t), t \geq 0\}$  defined in Section 2 is an  $\mathcal{F}$ -martingale with quadratic variation process given by*

$$\langle\langle \hat{R} \rangle\rangle_t = \sum_{i=1}^{N_A(\lambda t)} \int_0^{(t-\tau_i) \wedge W(\tau_i-) \wedge p_i} h(u) du.$$

*Proof.* We prove that  $\langle\langle \hat{R} \rangle\rangle_t = A_{\theta^{(N)}, \eta}^N(t)$  a.s., where the latter is defined by (5.32) in [15]. The result then follows from Lemma 5.4 in [15].

By (5.32) in [15]

$$A_{\theta^{(N)}, \eta}^N(t) = \int_0^t \left( \int_{[0, H^r)} \mathbb{1}_{[0, \chi^{(N)}(s-)}(x) h^r(x) \eta_s^{(N)}(dx) \right) ds, \quad (19)$$

where  $[0, H^r)$  and  $h^r$  are the support of the distribution and the hazard rate of the patience times,  $\eta_s^{(N)}$  is a measure that keep tracks of the time each customer, who already arrived at the system, spent in the system and whose waiting time has not reached his patience,  $\chi^{(N)}$  represents the waiting time if the head-of-line customer in the queue at time  $t$ .

First observe that for  $t \geq 0$

$$\mathbb{1}_{[0, \chi^{(N)}(t-)}(x) \eta_s^{(N)} = \sum_{i=1}^{N_A(\lambda t)} \delta_{t-\tau_i} \mathbb{1}_{\{t-\tau_i \leq W(\tau_i-)\}} \mathbb{1}_{\{p_i \geq t-\tau_i\}},$$

where  $\delta$  is the Dirac measure. Therefore,

$$\int_{[0, H^r)} \mathbb{1}_{[0, \chi^{(N)}(s-)}(x) h^r(x) \eta_s^{(N)}(dx) = \sum_{i=1}^{N_A(\lambda t)} h^r(s - \tau_i) \mathbb{1}_{\{s-\tau_i \leq W(\tau_i-)\}} \mathbb{1}_{\{p_i \geq s-\tau_i\}} \mathbb{1}_{\{\tau_i \leq s\}}.$$

This with (19) gives

$$\begin{aligned} A_{\theta^{(N)}, \eta}^N(t) &= \sum_{i=1}^{N_A(\lambda t)} \int_0^t h^r(s - \tau_i) \mathbb{1}_{\{s-\tau_i \leq W(\tau_i-)\}} \mathbb{1}_{\{p_i \geq s-\tau_i\}} \mathbb{1}_{\{\tau_i \leq s\}} ds \\ &= \sum_{i=1}^{N_A(\lambda t)} \int_{\tau_i}^{\tau_i + (W(\tau_i-) \wedge p_i) \wedge t - \tau_i} h^r(s) ds \\ &= \langle\langle \hat{R} \rangle\rangle_t, \end{aligned}$$

completing the proof.

We note that the filtration  $\mathcal{F}_t^{(N)}$  defined on page 11 of [15] is the same with the filtration  $\mathcal{F}$  here, since given all the processes there are recoverable from our processes (1)-(3) and vice versa. We note that the process  $s_j^{(N)}$  defined in [15] can easily be added to our filtration without altering our results.  $\square$

## 7 Fluid Limit

In this Section, we obtain our fluid limit results for the  $GI/M/n + GI$  queue. These results are helpful only in so far as they allow to obtain our diffusion limit result in the Section that follows. We now define the following fluid scaled quantities. For each  $t \geq 0$ , let

$$\begin{aligned}\bar{Q}_0^n &= \frac{Q_0^n}{n}, \\ \bar{Q}^n(t) &= \frac{Q^n(t)}{n}, \\ \bar{N}_A^n(t) &= \frac{\hat{N}_A(nt)}{n}, \\ \bar{N}_D^n(t) &= \frac{\hat{N}_D(nt)}{n},\end{aligned}$$

and

$$\bar{R}^n(t) = \frac{\hat{R}^n(t)}{n}.$$

It then follows dividing (18) through by  $n$  and noting that

$$\frac{\hat{N}_A(\lambda^n t)}{n} = \bar{N}_A^n(n^{-1}\lambda^n t),$$

and

$$\frac{N_D\left(\mu \int_0^t (Q^n(s) \wedge n) ds\right) - \mu \int_0^t (Q(s) \wedge n) ds}{n} = \bar{N}_D^n\left(\mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds\right),$$

that we obtain

$$\begin{aligned}\bar{Q}^n(t) &= \bar{Q}_0^n + \bar{N}_A^n(n^{-1}\lambda^n t) - \bar{N}_D^n\left(\mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds\right) - \bar{R}^n(t) + (n^{-1}\lambda^n - \mu)t \\ &\quad - \mu \int_0^t (\bar{Q}^n(s) - 1)^- ds - \frac{1}{n} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{(t-\tau_i/\lambda^n) \wedge W^n((\tau_i/\lambda^n)^-) \wedge p_i} h(u) du.\end{aligned}\tag{20}$$

Note that we use  $(\tau_i/\lambda^n)$  since this represents the arrival time of the  $i^{\text{th}}$  customer in the  $n^{\text{th}}$  system. The following is now the first main result of this Section.

**Proposition 7.1.** *If  $n^{-1}\lambda^n \rightarrow \lambda$  and  $\bar{Q}_0^n \Rightarrow \bar{Q}_0$  as  $n \rightarrow \infty$ , then*

$$\left( \bar{N}_A^n(\lambda^n e), \bar{N}_D^n \left( \mu^n \int_0^e (\bar{Q}^n(s) \wedge 1) ds \right), \bar{R}^n \right) \Rightarrow (0, 0, 0),$$

as  $n \rightarrow \infty$ .

*Proof.* By Theorem 3.9 of [2] it suffices to show that each term converges to zero on its own. First note that since by assumption  $\lambda^n e \rightarrow \lambda e$  and by the Functional Weak Law of Large Numbers [27]  $\bar{N}_A^n \Rightarrow 0$  as  $n \rightarrow \infty$ , it follows by the Random Rime Change Theorem [2] that  $\bar{N}_A^n(\lambda^n e) \Rightarrow 0$  as  $n \rightarrow \infty$ .

Next, note that for each  $t \geq 0$ ,  $Q(t) \leq Q_0 + N_A(\lambda t)$ , and hence for each  $T \geq 0$ , we have that

$$\sup_{0 \leq t \leq T} \left| \bar{N}_D^n \left( \mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) \right| \leq \sup_{0 \leq t \leq T} \left| \bar{N}_D^n \left( \mu \int_0^t (\bar{Q}_0^n + \bar{N}_A^n(n^{-1}\lambda^n s) + n^{-1}\lambda^n s) ds \right) \right| \quad (21)$$

However, since  $\bar{Q}_0^n + \bar{N}_A^n(\lambda^n e) + \lambda^n e \Rightarrow \bar{Q}_0 + \lambda e$  by the assumption of the Proposition and  $\bar{Q}_0^n + \bar{N}_A^n(\lambda^n e) + \lambda^n e$  is also an increasing function, it follows that

$$\mu \int_0^e (\bar{Q}_0^n + \bar{N}_A^n(n^{-1}\lambda^n s) + n^{-1}\lambda^n s) ds \Rightarrow \mu \int_0^e (\bar{Q}_0 + \lambda s) ds,$$

as  $n \rightarrow \infty$ . Hence, since by the Functional Weak Law of Large Numbers [2],  $\bar{N}_D^n \Rightarrow 0$  as  $n \rightarrow \infty$ , it follows by the Random Time Change Theorem [2] that  $\bar{N}_D^n (\bar{Q}_0^n + \bar{N}_A^n(\lambda^n e) + \lambda^n e) \Rightarrow 0$  as  $n \rightarrow \infty$ . Thus, by the Continuous Mapping Theorem [27],

$$\sup_{0 \leq t \leq T} \left| \bar{N}_D^n (\bar{Q}_0^n + \bar{N}_A^n(\lambda^n e) + \lambda^n e) \right| \Rightarrow 0,$$

as  $n \rightarrow \infty$  and hence, as a result of (21) above it follows that

$$\bar{N}_D^n \left( \mu^n \int_0^e (\bar{Q}^n(s) \wedge 1) ds \right) \Rightarrow 0,$$

as  $n \rightarrow \infty$ .

It remains to show that  $\bar{R}^n \Rightarrow 0$  as  $n \rightarrow \infty$ . By Proposition (6.1) we have that  $\bar{R}^n$  is a martingale with quadratic variation given by

$$\begin{aligned} \langle \langle \bar{R}^n \rangle \rangle_t &= \frac{1}{n^2} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{(t-\tau_i^n/\lambda^n) \wedge W^n(\tau_i^n/\lambda^n -) \wedge p_i} h(u) du \\ &\leq \frac{t}{n} \|h\|_\infty \frac{N_A(\lambda^n t)}{n}. \end{aligned}$$

Thus, for each  $T \geq 0$ , we have that

$$\sup_{0 \leq t \leq T} \langle \langle \bar{R}^n \rangle \rangle_t \leq \frac{T}{n} \|h\|_\infty \frac{N_A(\lambda^n T)}{n}. \quad (22)$$

However, since by the Functional Strong Law of Large Numbers [2], the assumption of the Proposition and the Random Time change Theorem we have that  $n^{-1}N_A(\lambda^n T) = n^{-1}N_A(n(n^{-1}\lambda^n)T) \Rightarrow \lambda T$  as  $n \rightarrow \infty$ , it follows by (22) that  $\langle \langle \bar{R}^n \rangle \rangle \Rightarrow 0$  as  $n \rightarrow \infty$  and hence by the Martingale Invariance Principle,  $\bar{R}^n \Rightarrow 0$  as  $n \rightarrow \infty$ .  $\square$

The following is now the main result of this Section.

**Theorem 7.2.** *If  $n^{-1}\lambda^n \rightarrow \lambda$  as  $n \rightarrow \infty$  where  $\lambda = \mu$  and  $\bar{Q}_0^n \Rightarrow 1$  as  $n \rightarrow \infty$ , then  $\bar{Q}^n \Rightarrow 1$  as  $n \rightarrow \infty$ .*

*Proof.* First note that

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{(t-\tau_i^n/\lambda^n) \wedge W^n(\tau_i^n/\lambda^n) \wedge p_i} h(s) ds \\
& \leq \frac{\|h\|_\infty}{n} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{(t-\tau_i^n/\lambda^n) \wedge W^n(\tau_i^n/\lambda^n) \wedge p_i} ds \\
& = \frac{\|h\|_\infty}{n} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^t \mathbf{1}\{0 \leq s - \tau_i/\lambda^n \leq W^n(\tau_i^n/\lambda^n) \wedge p_i\} ds \\
& = \frac{\|h\|_\infty}{n} \int_0^t \sum_{i=1}^{N_A(\lambda^n t)} \mathbf{1}\{0 \leq s - \tau_i/\lambda^n \leq W^n(\tau_i^n/\lambda^n) \wedge p_i\} ds \\
& = \|h\|_\infty \int_0^t (\bar{Q}^n(s) - 1)^+ ds.
\end{aligned}$$

Thus, by (18) and the triangle inequality, we have that

$$\begin{aligned}
|\bar{Q}^n(t) - 1| & \leq |\bar{Q}_0^n - 1| + |\bar{N}_A^n(n^{-1}\lambda^n t)| + \left| \bar{N}_D^n \left( \mu^n \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) \right| + |\bar{R}^n(t)| + |n^{-1}\lambda^n - \mu|t| \\
& \quad + \mu \int_0^t |\bar{Q}^n(s) - 1| ds + \|h\|_\infty \int_0^t |\bar{Q}^n(s) - 1| ds.
\end{aligned}$$

Moreover, for each  $T \geq r \geq 0$ ,

$$\begin{aligned}
& \sup_{0 \leq t \leq r} |\bar{Q}^n(t) - 1| \\
& \leq \sup_{0 \leq t \leq T} \left| |\bar{Q}_0^n - 1| + |\bar{N}_A^n(n^{-1}\lambda^n t)| + \left| \bar{N}_D^n \left( \mu^n \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) \right| + |\bar{R}^n(t)| + |n^{-1}\lambda^n - \mu|t| \right| \\
& \quad + (\mu + \|h\|_\infty) \int_0^r \sup_{0 \leq t \leq s} |\bar{Q}^n(t) - 1| ds.
\end{aligned}$$

It therefore follows by Gronwall's inequality [21] that for each  $T \geq 0$ ,

$$\begin{aligned}
& e^{-(\mu + \|h\|_\infty)T} \sup_{0 \leq t \leq T} |\bar{Q}^n(t) - 1| \tag{23} \\
& \leq \sup_{0 \leq t \leq T} \left| |\bar{Q}_0^n - 1| + |\bar{N}_A^n(n^{-1}\lambda^n t)| + \left| \bar{N}_D^n \left( \mu^n \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) \right| + |\bar{R}^n(t)| + |n^{-1}\lambda^n - \mu|t| \right|.
\end{aligned}$$

However since by the assumption of the Proposition, Proposition 7.1 and the Continuous Mapping Theorem [27] we have that

$$\sup_{0 \leq t \leq T} \left| |\bar{Q}_0^n - 1| + |\bar{N}_A^n(n^{-1}\lambda^n t)| + \left| \bar{N}_D^n \left( \mu^n \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) \right| + |\bar{R}^n(t)| + |n^{-1}\lambda^n - \mu|t| \right| \Rightarrow 0,$$

as  $n \rightarrow \infty$ , it follows by (23) that  $\bar{Q}^n \Rightarrow 1$  as  $n \rightarrow \infty$ , which completes the proof.  $\square$

## 8 Asymptotic Equivalence of $\tilde{Q}^{n,+}$ and $\tilde{W}^n$

In this Section, we prove an asymptotic equivalence between the limiting queue length process and the limiting virtual waiting time process. The heavy-traffic regime which we consider is identical to that in Section 7 with some additional assumptions. For each  $n \geq 1$  and  $t \geq 0$  let us first define the diffusion scaled quantities

$$\begin{aligned}\tilde{N}_A(t) &= \frac{N_A(nt) - nt}{\sqrt{n}}, \\ \tilde{N}_D(t) &= \frac{N_D(nt) - nt}{\sqrt{n}}\end{aligned}$$

and

$$\tilde{R}^n(t) = \frac{R^n(t)}{\sqrt{n}}.$$

It then follows from (18) after a little bit of algebra that

$$\begin{aligned}\tilde{Q}^n(t) &= \tilde{Q}_0^n + \tilde{N}_A(n^{-1}\lambda^n t) - \tilde{N}_D\left(\mu \int_0^t (\tilde{Q}^n(s) \wedge 1) ds\right) + \tilde{R}^n(t) + n^{1/2}(n^{-1}\lambda^n - \mu)t \\ &\quad - \mu \int_0^t \tilde{Q}^{n,-}(s) ds - \frac{1}{\sqrt{n}} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{(t-\tau_i/\lambda^n) \wedge W^n(\tau_i/\lambda^n) \wedge p_i^n} h(u) du.\end{aligned}\quad (24)$$

The following result is now the first main result of this Section. It follows from our fluid limit results in Section 7.

**Proposition 8.1.** *If  $n^{-1}\lambda^n \rightarrow \lambda$  as  $n \rightarrow \infty$  where  $\lambda = \mu$  and  $\tilde{Q}_0^n \Rightarrow \tilde{Q}_0^n$  as  $n \rightarrow \infty$ , then*

$$\left(\tilde{Q}_0^n, \tilde{N}_A^n(n^{-1}\lambda^n e), \tilde{N}_D^n\left(\mu \int_0^e (\tilde{Q}^n(s) \wedge 1) ds\right), \tilde{R}^n\right) \Rightarrow (\tilde{Q}_0, \sigma_A \tilde{B}_1(\lambda e), \tilde{B}_2(\lambda e), 0),$$

as  $n \rightarrow \infty$ , where  $\tilde{B}_1$  and  $\tilde{B}_2$  are independent, standard Brownian motions both of which are independent of  $\tilde{Q}_0^n$

*Proof.* By the Functional Central Limit Theorem for renewal process [2] and the assumed independence of  $Q_0, N_A$  and  $N_D$ , it follows that  $(\tilde{Q}_0^n, \tilde{N}_A^n, \tilde{N}_D^n) \Rightarrow (\tilde{Q}_0, \sigma_A \tilde{B}_1, \tilde{B}_2)$  as  $n \rightarrow \infty$ . By assumption, we have that  $n^{-1}\lambda^n \Rightarrow e$  as  $n \rightarrow \infty$ . Also, by the assumption of the Proposition and Theorem 7.2 we have that  $\tilde{Q}^n \Rightarrow 1$  as  $n \rightarrow \infty$ . Hence, by the Bounded Convergence Theorem [10],

$$\mu \int_0^e (\tilde{Q}^n(s) \wedge 1) ds \Rightarrow \mu e,$$

as  $n \rightarrow \infty$ . It therefore follows by the Random Time Change Theorem [2] that

$$\left(\tilde{Q}_0^n, \tilde{N}_A^n(n^{-1}\lambda^n e), \tilde{N}_D^n\left(\mu \int_0^e (\tilde{Q}^n(s) \wedge 1) ds\right)\right) \Rightarrow (\tilde{Q}_0, \sigma_A \tilde{B}_1(\lambda e), \tilde{B}_2(\mu e)),$$

as  $n \rightarrow \infty$ .

By Theorem 3.9 of [2], it now remains to show that  $\tilde{R}^n \Rightarrow 0$  as  $n \rightarrow \infty$ . By Proposition 6.1 of Section 6.2,  $\tilde{R}^n$  is a martingale with quadratic variation

$$\langle \langle \tilde{R}^n \rangle \rangle_t = \frac{1}{n} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{(t-\tau_i^n/\lambda^n) \wedge W^n(\tau_i^n/\lambda^n) \wedge p_i^n} h(u) du.$$

However, as was already demonstrated in the proof of Theorem 7.2,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{(t-\tau_i^n) \wedge W^n(\tau_i^n) \wedge p_i} h(s) ds &\leq \|h\|_\infty \int_0^t (\bar{Q}^n(s) - 1)^+ ds \\ &\Rightarrow 0, \end{aligned}$$

where the final convergence follows from the assumptions of the Proposition and Theorem 7.2. It therefore follows from the Martingale Invariance Principle [9] that  $\tilde{R}^n \Rightarrow 0$  as  $n \rightarrow \infty$ , which completes the proof.  $\square$

We now have the following Proposition which prepares for us the statement of the last result of this Section.

**Proposition 8.2.** *If  $n^{-1}\lambda^n \rightarrow \lambda$  as  $n \rightarrow \infty$  where  $\lambda = \mu$  and  $\tilde{Q}_0^n \Rightarrow \tilde{Q}_0$  as  $n \rightarrow \infty$  and  $n^{1/2}(n^{-1}\lambda^n - \mu) \rightarrow -\beta\mu$  as  $n \rightarrow \infty$ , then  $\{\tilde{Q}^n\}$  is tight.*

*Proof.* In order to show that  $\{\tilde{Q}^n\}$  is tight we must verify that conditions (i) and (ii) of Theorem 13.2 of [2] are satisfied. We begin with condition (i). We must show that for each  $T > 0$  and  $\varepsilon > 0$  there exists a  $K_\varepsilon^T > 0$  such that  $\mathbb{P}(\sup_{0 \leq t \leq T} |\tilde{Q}_t^n| > K_\varepsilon^T) < \varepsilon$  for  $n \geq 1$ . First note that since, as in the proof of Theorem 7.2 we have that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{(t-\tau_i^n/\lambda^n) \wedge W^n(\tau_i^n/\lambda^n) \wedge p_i^n} h(s) ds \leq \|h\|_\infty \int_0^t \tilde{Q}^{n,+}(s) ds,$$

it follows from (24) that

$$\begin{aligned} \tilde{Q}^n(t) &\leq \tilde{Q}_0^n + \tilde{N}_A(n^{-1}\lambda^n t) - \tilde{N}_D \left( \mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) + \tilde{R}^n(t) + n^{1/2}(n^{-1}\lambda^n - \mu)t \\ &\quad - \mu \int_0^t \tilde{Q}^{n,-}(s) ds + \|h\|_\infty \int_0^t \tilde{Q}^{n,+}(s) ds. \end{aligned} \quad (25)$$

Taking supremums on both sides of (25), we therefore obtain that for each  $t \geq 0$ ,

$$\begin{aligned} &\sup_{0 \leq s \leq t} |\tilde{Q}^n(s)| \\ &\leq \sup_{0 \leq s \leq t} \left| \tilde{Q}_0^n + \tilde{N}_A(n^{-1}\lambda^n t) - \tilde{N}_D \left( \mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) + \tilde{R}^n(t) + n^{1/2}(n^{-1}\lambda^n - \mu)t \right| \\ &\quad + \mu \int_0^t \sup_{0 \leq u \leq s} |\tilde{Q}^{n,-}(u)| ds + \|h\|_\infty \int_0^t \sup_{0 \leq u \leq s} |\tilde{Q}^{n,+}(u)| ds, \end{aligned}$$

thus implying that for each  $T \geq 0$  and  $0 \leq t \leq T$ ,

$$\begin{aligned} & \sup_{0 \leq s \leq t} |\tilde{Q}^n(s)| \\ \leq & \sup_{0 \leq t \leq T} \left| \tilde{Q}_0^n + \tilde{N}_A(n^{-1}\lambda^n t) - \tilde{N}_D \left( \mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) + \tilde{R}^n(t) + n^{1/2}(n^{-1}\lambda^n - \mu)t \right| \\ & + (\mu + \|h\|_\infty) \int_0^t \sup_{0 \leq u \leq s} |\tilde{Q}^n(u)| ds. \end{aligned}$$

Hence, by Gronwall's inequality [27], we obtain that

$$\begin{aligned} & e^{-(\mu + \|h\|_\infty)T} \sup_{0 \leq t \leq T} |\tilde{Q}^n(t)| \tag{26} \\ \leq & \sup_{0 \leq t \leq T} \left| \tilde{Q}_0^n + \tilde{N}_A(n^{-1}\lambda^n t) - \tilde{N}_D \left( \mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) + \tilde{R}^n(t) + n^{1/2}(n^{-1}\lambda^n - \mu)t \right|. \end{aligned}$$

By Proposition 8.1 and the Continuous Mapping Theorem [27] we have that

$$\begin{aligned} & \tilde{Q}_0^n + \tilde{N}_A(n^{-1}\lambda^n e) - \tilde{N}_D \left( \mu \int_0^e (\bar{Q}^n(s) \wedge 1) ds \right) + \tilde{R}^n(e) + n^{1/2}(n^{-1}\lambda^n - \mu)e \\ \Rightarrow & \tilde{Q}_0 + \tilde{B}_1(\lambda e) - \tilde{B}_2(\lambda e) - \beta \mu e, \end{aligned}$$

as  $n \rightarrow \infty$ . Thus, by the only if portion of Theorem 13.2 of [2] we have condition (i) holds for

$$\left\{ \tilde{Q}_0^n + \tilde{N}_A(n^{-1}\lambda^n e) - \tilde{N}_D \left( \mu \int_0^e (\bar{Q}^n(s) \wedge 1) ds \right) + \tilde{R}^n(e) + n^{1/2}(n^{-1}\lambda^n - \mu)e \right\}$$

so condition (i) holds for  $\{\tilde{Q}^n\}$  as well as a result of (26).

We now verify that condition (ii) is satisfied. The proof follows similarly to the proof of condition (i) above. By (24) it follows that for each  $\delta \geq 0$ ,

$$\begin{aligned} & \tilde{Q}^n(t + \delta) - \tilde{Q}^n(t) \tag{27} \\ = & (\tilde{N}_A(n^{-1}\lambda^n(t + \delta)) - \tilde{N}_A(n^{-1}\lambda^n t)) \\ & - \left( \tilde{N}_D \left( \mu \int_0^{t+\delta} (\bar{Q}^n(s) \wedge 1) ds \right) - \tilde{N}_D \left( \mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) \right) + (\tilde{R}^n(t + \delta) - \tilde{R}^n(t)) \\ & + n^{1/2}(n^{-1}\lambda^n - \mu)\delta - \mu \int_t^{t+\delta} \tilde{Q}^{n,-}(s) ds \\ & - \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{N_A(\lambda^n(t+\delta))} \int_0^{(t+\delta-\tau_i/\lambda^n) \wedge W^n(\tau_i/\lambda^n-)} h(u) du \right. \\ & \left. - \frac{1}{\sqrt{n}} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{(t-\tau_i/\lambda^n) \wedge W^n(\tau_i/\lambda^n-)} h(u) du \right). \end{aligned}$$

Moreover, as in the proof of Theorem 7.2 it may be shown that

$$\left( \frac{1}{\sqrt{n}} \sum_{i=1}^{N_A(\lambda^n(t+\delta))} \int_0^{(t+\delta-\tau_i/\lambda^n) \wedge W^n(\tau_i/\lambda^n-)} h(u) du - \frac{1}{\sqrt{n}} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{(t-\tau_i/\lambda^n) \wedge W^n(\tau_i/\lambda^n-)} h(u) du \right)$$

$$\leq \|h\|_\infty \int_t^{t+\delta} \tilde{Q}^{n,+}(s) ds.$$

Next, note that

$$\|h\|_\infty \int_t^{t+\delta} \tilde{Q}^{n,+}(s) ds \leq \|h\|_\infty \delta \sup_{0 \leq t \leq T} |\tilde{Q}^n(s)|$$

and

$$\mu \int_t^{t+\delta} \tilde{Q}^{n,-}(s) ds \leq \mu \delta \sup_{0 \leq t \leq T} |\tilde{Q}^n(s)|.$$

It therefore follows combining the above with (27) we have that for each  $T \geq 0$ ,

$$\begin{aligned} & \sup_{0 \leq t \leq T} |\tilde{Q}^n(t+\delta) - \tilde{Q}^n(t)| \\ & \leq \sup_{0 \leq t \leq T} |\tilde{N}_A(\lambda(t+\delta)) - \tilde{N}_A(\lambda t)| \\ & \quad + \sup_{0 \leq t \leq T} \left| \tilde{N}_D \left( \mu \int_0^{t+\delta} (\tilde{Q}^n(s) \wedge 1) ds \right) - \tilde{N}_D \left( \mu \int_0^t (\tilde{Q}^n(s) \wedge 1) ds \right) \right| \\ & \quad + \sup_{0 \leq t \leq T} |\tilde{R}^n(t+\delta) - \tilde{R}^n(t)| + n^{1/2}(n^{-1}\lambda^n - \mu)\delta + \delta(\|h\|_\infty + \mu) \sup_{0 \leq t \leq T} |\tilde{Q}^n(s)|. \end{aligned}$$

Thus, condition (ii) of Theorem 13.2 of [2] is now seen to be satisfied by virtue of (i) above and Proposition 8.1 in conjunction with the only if portion of Theorem 13.2 of [2].  $\square$

The following result is now the main result of this Section. Let us define the diffusion scaled virtual waiting time at time for each  $n \geq 1$  and  $t \geq 0$ ,

$$\tilde{W}^n = \sqrt{n}W^n(t).$$

We then have the following.

**Proposition 8.3.** *If  $n^{-1}\lambda^n \rightarrow \lambda$  as  $n \rightarrow \infty$  where  $\lambda = \mu$  and  $\tilde{Q}_0^n \Rightarrow \tilde{Q}_0$  as  $n \rightarrow \infty$  and  $n^{1/2}(n^{-1}\lambda^n - \mu) \rightarrow -\beta\mu$  as  $n \rightarrow \infty$ , then  $|\tilde{Q}^{n,+} - \lambda\tilde{W}^n| \Rightarrow 0$  as  $n \rightarrow \infty$ .*

*Proof.* First recall by (3) that the virtual waiting time at time  $t \geq 0$  is given by

$$\begin{aligned} & W^n(t) \\ & = \inf \left\{ u \geq 0 : n^{-1/2} \left( N_D \left( \mu \int_0^{t+u} (Q^n(s) \wedge n) ds \right) - N_D \left( \mu \int_0^t (Q^n(s) \wedge n) ds \right) \right) \right. \\ & \quad + n^{-1/2} \left( \sum_{i=1}^{N_A(\lambda^n t)} 1\{p_i^n \leq W^n(\tau_i/\lambda^n)\} 1\{p_i^n \leq (t+u) - \tau_i/\lambda^n\} \right. \\ & \quad \left. \left. - \sum_{i=1}^{N_A(\lambda^n t)} 1\{p_i^n \leq W^n(\tau_i/\lambda^n)\} 1\{p_i^n \leq t - \tau_i/\lambda^n\} \right) \geq \tilde{Q}^{n,+}(t) \right\}. \end{aligned}$$

This may be rewritten as

$$\begin{aligned}
& \tilde{W}^n(t) \\
&= \inf \left\{ u \geq 0 : n^{-1/2} \left( N_D \left( \mu \int_0^{t+n^{-1/2}u} (Q^n(s) \wedge n) ds \right) - N_D \left( \mu \int_0^t (Q^n(s) \wedge n) ds \right) \right) \right. \\
&\quad + n^{-1/2} \left( \sum_{i=1}^{N_A(\lambda^n t)} 1\{p_i^n \leq W^n(\tau_i/\lambda^n)\} 1\{p_i^n \leq (t+n^{-1/2}u) - \tau_i/\lambda^n\} \right. \\
&\quad \left. \left. - \sum_{i=1}^{N_A(\lambda^n t)} 1\{p_i^n \leq W^n(\tau_i/\lambda^n)\} 1\{p_i^n \leq t - \tau_i/\lambda^n\} \right) \geq \tilde{Q}^{n,+}(t) \right\} \\
&= \inf \left\{ u \geq 0 : \mu u + \left( n^{-1/2} \left( N_D \left( \mu \int_0^{t+n^{-1/2}u} (Q^n(s) \wedge n) ds \right) - N_D \left( \mu \int_0^t (Q^n(s) \wedge n) ds \right) \right) - \mu u \right) \right. \\
&\quad + n^{-1/2} \left( \sum_{i=1}^{N_A(\lambda^n t)} 1\{p_i^n \leq W^n(\tau_i/\lambda^n)\} 1\{p_i^n \leq (t+n^{-1/2}u) - \tau_i/\lambda^n\} \right. \\
&\quad \left. \left. - \sum_{i=1}^{N_A(\lambda^n t)} 1\{p_i^n \leq W^n(\tau_i/\lambda^n)\} 1\{p_i^n \leq t - \tau_i/\lambda^n\} \right) \geq \tilde{Q}^{n,+}(t) \right\}. \tag{28}
\end{aligned}$$

We now make the claim that for each  $U \geq 0$  and  $T \geq 0$ ,

$$\sup_{0 \leq u \leq U} \sup_{0 \leq t \leq T} \left| n^{-1/2} \left( N_D \left( \mu \int_0^{t+n^{-1/2}u} (Q^n(s) \wedge n) ds \right) - N_D \left( \mu \int_0^t (Q^n(s) \wedge n) ds \right) \right) - \mu u \right| \Rightarrow 0,$$

as  $n \rightarrow \infty$ . In order to see this, first note that for each  $t \geq 0$  and  $u \geq 0$ ,

$$\begin{aligned}
& \left( n^{-1/2} \left( N_D \left( \mu \int_0^{t+n^{-1/2}u} (Q^n(s) \wedge n) ds \right) - N_D \left( \mu \int_0^t (Q^n(s) \wedge n) ds \right) \right) - \mu u \right) \\
&= n^{-1/2} \left( N_D \left( n\mu \int_0^{t+n^{-1/2}u} (\bar{Q}^n(s) \wedge 1) ds \right) - n\mu \int_0^{t+n^{-1/2}u} (\bar{Q}^n(s) \wedge 1) ds \right) \\
&\quad - n^{-1/2} \left( N_D \left( n\mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) - n\mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) \\
&\quad + \left( n^{1/2} \mu \int_t^{t+n^{-1/2}u} (\bar{Q}^n(s) \wedge 1) ds - \mu u \right).
\end{aligned}$$

By the assumptions of the Proposition and Theorem 7.2 it is clear that

$$\sup_{0 \leq u \leq U} \sup_{0 \leq t \leq T} \left| n^{1/2} \mu \int_t^{t+n^{-1/2}u} (\bar{Q}^n(s) \wedge 1) ds - \mu u \right| \Rightarrow 0,$$

Next, note that

$$n^{-1/2} \left( N_D \left( n\mu \int_0^{t+n^{-1/2}u} (\bar{Q}^n(s) \wedge 1) ds \right) - n\mu \int_0^{t+n^{-1/2}u} (\bar{Q}^n(s) \wedge 1) ds \right)$$

$$\begin{aligned}
& -n^{-1/2} \left( N_D \left( n\mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) - n\mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) \\
&= \tilde{N}_D^n \left( \mu \int_0^{t+n^{-1/2}u} (\bar{Q}^n(s) \wedge 1) ds \right) - \tilde{N}_D^n \left( \mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right).
\end{aligned}$$

It therefore follows that

$$\begin{aligned}
& \sup_{0 \leq u \leq U} \sup_{0 \leq t \leq T} \left| n^{-1/2} \left( N_D \left( n\mu \int_0^{t+n^{-1/2}u} (\bar{Q}^n(s) \wedge 1) ds \right) - n\mu \int_0^{t+n^{-1/2}u} (\bar{Q}^n(s) \wedge 1) ds \right) \right. \\
& \quad \left. - n^{-1/2} \left( N_D \left( n\mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) - n\mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) \right| \\
&= \sup_{0 \leq u \leq U} \sup_{0 \leq t \leq T} \left| \tilde{N}_D^n \left( \mu \int_0^{t+n^{-1/2}u} (\bar{Q}^n(s) \wedge 1) ds \right) - \tilde{N}_D^n \left( \mu \int_0^t (\bar{Q}^n(s) \wedge 1) ds \right) \right| \\
&\leq \sup_{0 \leq u \leq U} \sup_{0 \leq t \leq \mu T} \left| \tilde{N}_D^n \left( t + n^{-1/2}u \right) - \tilde{N}_D^n \left( t \right) \right| \\
&\Rightarrow 0,
\end{aligned}$$

as  $n \rightarrow \infty$ , where the final convergence follows since by the Functional Central Limit Theorem for renewal processes [2],  $\tilde{N}_D^n \Rightarrow \tilde{B}$  as  $n \rightarrow \infty$ , where  $\tilde{B}$  is a standard Brownian motion with  $\mathbb{P}$ -a.s. continuous sample paths. Thus, the claim is proven.

Next, we claim that

$$\begin{aligned}
& \sup_{0 \leq u \leq U} \sup_{0 \leq t \leq T} \left| n^{-1/2} \left( \sum_{i=1}^{N_A(\lambda^n t)} 1\{p_i^n \leq W^n(\tau_i/\lambda^n)\} 1\{p_i^n \leq (t + n^{-1/2}u) - \tau_i/\lambda^n\} \right. \right. \\
& \quad \left. \left. - \sum_{i=1}^{N_A(\lambda^n t)} 1\{p_i^n \leq W^n(\tau_i/\lambda^n)\} 1\{p_i^n \leq t - \tau_i/\lambda^n\} \right) \right| \\
&\Rightarrow 0,
\end{aligned}$$

as  $n \rightarrow \infty$ . First note that

$$\begin{aligned}
& n^{-1/2} \left( \sum_{i=1}^{N_A(\lambda^n t)} 1\{p_i^n \leq W^n(\tau_i/\lambda^n)\} 1\{p_i^n \leq (t + n^{-1/2}u) - \tau_i/\lambda^n\} \right. \\
& \quad \left. - \sum_{i=1}^{N_A(\lambda^n t)} 1\{p_i^n \leq W^n(\tau_i/\lambda^n)\} 1\{p_i^n \leq t - \tau_i/\lambda^n\} \right) \\
&\leq n^{-1/2} \left( \sum_{i=1}^{N_A(\lambda^n(t+n^{-1/2}u))} 1\{p_i^n \leq W^n(\tau_i/\lambda^n)\} 1\{p_i^n \leq (t + n^{-1/2}u) - \tau_i/\lambda^n\} \right. \\
& \quad \left. - \sum_{i=1}^{N_A(\lambda^n t)} 1\{p_i^n \leq W^n(\tau_i/\lambda^n)\} 1\{p_i^n \leq t - \tau_i/\lambda^n\} \right)
\end{aligned}$$

$$\begin{aligned}
&= \tilde{R}^n(t + n^{-1/2}u) - \tilde{R}^n(t) + n^{-1/2} \left( \sum_{i=1}^{N_A(\lambda^n(t+n^{-1/2}u))} \int_0^{(t+n^{-1/2}u-\tau_i/\lambda^n)\wedge W(\tau_i/\lambda^n-)\wedge p_i^n} h(u)du \right. \\
&\quad \left. - \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{(t-\tau_i/\lambda^n)\wedge W(\tau_i/\lambda^n-)\wedge p_i^n} h(u)du \right).
\end{aligned}$$

By the assumptions of the Proposition and Proposition 8.1.  $\tilde{R}^n \Rightarrow 0$  as  $n \rightarrow \infty$  and hence it is clear that

$$\sup_{0 \leq u \leq U} \sup_{0 \leq t \leq T} |\tilde{R}^n(t + n^{-1/2}u) + \tilde{R}^n(t)| \Rightarrow 0,$$

as  $n \rightarrow \infty$ . Moreover, as in the proof of Theorem 7.2, it follows by the assumptions of the Proposition and Theorem 7.2

$$\begin{aligned}
&\sup_{0 \leq u \leq U} \sup_{0 \leq t \leq T} n^{-1/2} \left| \sum_{i=1}^{N_A(\lambda^n(t+n^{-1/2}u))} \int_0^{(t+n^{-1/2}u-\tau_i/\lambda^n)\wedge W(\tau_i/\lambda^n-)\wedge p_i^n} h(u)du \right. \\
&\quad \left. - \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{(t-\tau_i/\lambda^n)\wedge W(\tau_i/\lambda^n-)\wedge p_i^n} h(u)du \right| \\
&\leq \sup_{0 \leq u \leq U} \sup_{0 \leq t \leq T} n^{-1/2} \|h\|_\infty \int_t^{t+n^{-1/2}u} (Q^n(s) - n)^+ ds \\
&= \sup_{0 \leq u \leq U} \sup_{0 \leq t \leq T} \|h\|_\infty \int_t^{t+n^{-1/2}u} \tilde{Q}^{n,+}(s) ds \\
&\leq U \|h\|_\infty \sup_{0 \leq t \leq T+n^{-1/2}U} (\bar{Q}^n(u) - 1)^+ \\
&\Rightarrow 0,
\end{aligned}$$

as  $n \rightarrow \infty$ . Thus, by the triangle inequality, the claim is proven.

Now note that Proposition 8.2 we have that  $\{\tilde{Q}^n\}$  is stochastically bounded. Thus, for each  $T > 0$  and  $\varepsilon > 0$  there exists a  $K_\varepsilon^T > 0$  such that  $\mathbb{P}(\sup_{0 \leq t \leq T} |\tilde{Q}_t^n| > K_\varepsilon^T) < \varepsilon$  for  $n \geq 1$ . Thus, by the representation 28 for  $\tilde{W}^n$  and the previous two results it is easy to see that for each  $T \geq 0$ ,  $\sup_{0 \leq t \leq T} |\tilde{W}^n(t) - \mu^{-1} \tilde{Q}^{n,+}(t)| \Rightarrow 0$  as  $n \rightarrow \infty$  and hence the claim of the Proposition is proven as a result of the fact that  $\lambda = \mu$ . □

## 9 Heavy-Traffic Limit

In this Section, we obtain our main result of the paper, Theorem 3.1. For each  $t \geq 0$ , let us first define the process

$$B(t) = R(t) - \sum_{i=1}^{N_A(\lambda t)} F(W(\tau_i-)).$$

Moreover, for each  $n \geq 1$ , let

$$\tilde{B}^n(t) = \frac{B^n(t)}{\sqrt{n}}.$$

We then have in a similar derivation leading up to system equation (18) that

$$\begin{aligned} \tilde{Q}^n(t) &= \tilde{Q}_0^n + \tilde{N}_A(n^{-1}\lambda^n t) - \tilde{N}_D \left( \mu \int_0^t (\tilde{Q}^n(s) \wedge 1) ds \right) + \tilde{B}^n(t) + n^{1/2}(n^{-1}\lambda^n - \mu)t \\ &\quad - \mu \int_0^t \tilde{Q}^{n,-}(s) ds - \frac{1}{\sqrt{n}} \sum_{i=1}^{N_A(\lambda^n t)} F^n(W^n(\tau_i/\lambda^n-)). \end{aligned} \quad (29)$$

Moreover, letting

$$\tilde{\delta}^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{N_A(\lambda^n t)} F^n(W^n(\tau_i/\lambda^n-)) - \lambda \int_0^t \int_0^{\tilde{Q}^{n,+}(s)} h(u) du ds,$$

it follows that

$$\begin{aligned} \tilde{Q}^n(t) &= \tilde{Q}_0^n + \tilde{N}_A(n^{-1}\lambda^n t) - \tilde{N}_D \left( \mu \int_0^t (\tilde{Q}^n(s) \wedge 1) ds \right) + \tilde{B}^n(t) + \tilde{\delta}^n(t) \\ &\quad + n^{1/2}(n^{-1}\lambda^n - \mu)t - \mu \int_0^t \tilde{Q}^{n,-}(s) ds - \lambda \int_0^t \int_0^{\tilde{Q}^{n,+}(s)} h(u) du ds. \end{aligned} \quad (30)$$

The following is now our first result of the Section.

**Proposition 9.1.** *If  $n^{-1}\lambda^n \rightarrow \lambda$  as  $n \rightarrow \infty$  where  $\lambda = \mu$  and  $\tilde{Q}_0^n \Rightarrow \tilde{Q}_0$  as  $n \rightarrow \infty$  and  $n^{1/2}(n^{-1}\lambda^n - \mu) \rightarrow -\beta\mu$  as  $n \rightarrow \infty$ , then  $\tilde{B}^n \Rightarrow 0$  as  $n \rightarrow \infty$ .*

*Proof.* First note that  $\tilde{B}^n$  is a martingale with quadratic variation

$$\langle \tilde{B}^n \rangle_t = \frac{1}{n} \sum_{i=1}^{N_A(\lambda^n t)} F^n(W^n(\tau_i^n-)).$$

Now, by Proposition 8.2 we have that  $\{\tilde{Q}^n\}$  is stochastically bounded and hence by Proposition 8.3 it follows that  $\{\tilde{W}^n\}$  is stochastically bounded as well. Let  $K_\varepsilon^T$  be such that  $\mathbb{P}(\sup_{0 \leq t \leq T} |\tilde{W}^n(t)| > K_\varepsilon^T) < \varepsilon$  for  $n \geq 1$ . It then follows that for each  $\delta \geq 0$ ,

$$\begin{aligned} \mathbb{P} \left( \frac{1}{n} \sum_{i=1}^{N_A(\lambda^n T)} F^n(W^n(\tau_i^n-)) \geq \delta \right) &\leq \mathbb{P} \left( \sup_{0 \leq t \leq T} |\tilde{W}^n(t)| > K_\varepsilon^T \right) \\ &\quad + \mathbb{P} \left( \frac{1}{n} N_A(\lambda^n T) F^n(n^{-1/2} K_\varepsilon^T) \geq \delta \right) \\ &\leq \varepsilon + \mathbb{P} \left( \frac{1}{n} N_A(\lambda^n T) F^n(n^{-1/2} K_\varepsilon^T) \geq \delta \right). \end{aligned}$$

However, since by the Functional Strong Law of Large Numbers [2], the fact that  $n^{-1}\lambda^n \rightarrow \lambda$  and the Random Time Change Theorem [2] we have that  $n^{-1}N(\lambda^n T) = n^{-1}N(n(n^{-1}\lambda^n)T) \Rightarrow \lambda T$  and also

$$\begin{aligned}
F^n(n^{-1/2}K_\varepsilon^T) &= 1 - \exp\left(\int_0^{n^{-1/2}K_\varepsilon^T} h^n(u)du\right) \\
&= 1 - \exp\left(-\frac{1}{\sqrt{n}}\int_0^{K_\varepsilon^T} h(u)du\right) \\
&\leq 1 - \exp\left(-\frac{1}{\sqrt{n}}\|h\|_\infty K_\varepsilon^T\right) \\
&\rightarrow 0,
\end{aligned}$$

as  $n \rightarrow \infty$ , it follows that

$$\limsup_{n \rightarrow \infty} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{N_A(\lambda^n T)} F^n(W^n(\tau_i^n -)) \geq \delta\right) \leq \varepsilon.$$

Thus, since  $\delta$  and  $\varepsilon$  were arbitrary and  $\langle \tilde{B}^n \rangle$  is an increasing process, it follows that  $\langle \tilde{B}^n \rangle \Rightarrow 0$  as  $n \rightarrow \infty$ . The result now follows by the Martingale Invariance Principle [9].  $\square$

We next have the following which is our final result before the main result of the Section.

**Proposition 9.2.** *If  $n^{-1}\lambda^n \rightarrow \lambda$  as  $n \rightarrow \infty$  where  $\lambda = \mu$  and  $\tilde{Q}_0^n \Rightarrow \tilde{Q}_0$  as  $n \rightarrow \infty$  and  $n^{1/2}(n^{-1}\lambda^n - \mu) \rightarrow -\beta\mu$  as  $n \rightarrow \infty$ , then  $\tilde{\delta}^n \Rightarrow 0$  as  $n \rightarrow \infty$ .*

*Proof.* First note that

$$\begin{aligned}
\tilde{\delta}^n(t) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{N_A(\lambda^n t)} F^n(W^n(\tau_i/\lambda^n -)) - \int_0^t \int_0^{\tilde{Q}^{n,+}(s)} h(u)duds \\
&= \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{N_A(\lambda^n t)} F^n(W^n(\tau_i/\lambda^n -)) - \frac{1}{n} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{\tilde{W}^n(\tau_i/\lambda^n -)} h(u)du \right) \\
&\quad + \left( \frac{1}{n} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{\tilde{W}^n(\tau_i/\lambda^n -)} h(u)du - \lambda \int_0^t \int_0^{\tilde{W}^n(s)} h(u)duds \right) \\
&\quad + \lambda \left( \int_0^t \int_0^{\tilde{W}^n(s)} h(u)duds - \int_0^t \int_0^{\tilde{Q}^{n,+}(s)} h(u)duds \right).
\end{aligned}$$

We now claim that each of the terms in the parenthesis above converges weakly to 0 as  $n$  goes to  $\infty$ .

First note that

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i=1}^{N_A(\lambda^n t)} F^n(W^n(\tau_i/\lambda^n-)) - \frac{1}{n} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{\tilde{W}^n(\tau_i/\lambda^n-)} h(u) du \\
&= \frac{1}{n} \sum_{i=1}^{N_A(\lambda^n t)} \left( \sqrt{n} F^n(W^n(\tau_i/\lambda^n-)) - \int_0^{\tilde{W}^n(\tau_i/\lambda^n-)} h(u) du \right) \\
&\leq \frac{1}{n} N_A(\lambda^n T) \sup_{0 \leq t \leq T} \left( \sqrt{n} F^n(W^n(t)) - \int_0^{\tilde{W}^n(t)} h(u) du \right).
\end{aligned} \tag{31}$$

Now note that by Taylor's Theorem [8] for each  $t \geq 0$ ,

$$\begin{aligned}
\sqrt{n} F^n(W^n(t)) &= \sqrt{n} \left( 1 - \exp \left( - \int_0^{W^n(t)} h^n(u) du \right) \right) \\
&= \sqrt{n} \left( 1 - \exp \left( - \frac{1}{\sqrt{n}} \int_0^{\tilde{W}^n(t)} h(u) du \right) \right) \\
&= \int_0^{\tilde{W}^n(t)} h(u) du + \sqrt{n} \int_0^{\frac{1}{\sqrt{n}} \int_0^{\tilde{W}^n(t)} h(u) du} \frac{1}{2} e^{-t} \left( \frac{1}{\sqrt{n}} \int_0^{\tilde{W}^n(t)} h(u) du - t \right) dt,
\end{aligned}$$

and hence by (32), for each  $T \geq 0$ ,

$$\begin{aligned}
& \sup_{0 \leq t \leq T} \frac{1}{\sqrt{n}} \sum_{i=1}^{N_A(\lambda^n t)} F^n(W^n(\tau_i/\lambda^n-)) - \frac{1}{n} \sum_{i=1}^{N_A(\lambda^n t)} \int_0^{\tilde{W}^n(\tau_i/\lambda^n-)} h(u) du \\
&\leq \frac{1}{n} N_A(\lambda^n T) \sup_{0 \leq t \leq T} \left( \sqrt{n} \int_0^{\frac{1}{\sqrt{n}} \int_0^{\tilde{W}^n(t)} h(u) du} \frac{1}{2} e^{-t} \left( \frac{1}{\sqrt{n}} \int_0^{\tilde{W}^n(t)} h(u) du - t \right) dt \right) \\
&\leq \frac{1}{n} N_A(\lambda^n T) \frac{1}{2} \sup_{0 \leq t \leq T} \sqrt{n} \left( \frac{1}{\sqrt{n}} \int_0^{\tilde{W}^n(t)} h(u) du \right)^2 \\
&\leq \frac{\|h\|_\infty}{2\sqrt{n}} \left( \frac{1}{n} N_A(\lambda^n T) \right) \left( \sup_{0 \leq t \leq T} (\tilde{W}^n(t))^2 \right).
\end{aligned}$$

However, since as in the proof of Proposition 9.1,  $n^{-1} N_A(\lambda^n T) \Rightarrow \lambda T$  as  $n \rightarrow \infty$  and since by Proposition 8.2 we have that  $\{\sup_{0 \leq t \leq T} (\tilde{W}^n(t))^2\}$  is stochastically bounded, it follows by Slutsky's Theorem [9] that

$$\frac{\|h\|_\infty}{2\sqrt{n}} \left( \frac{1}{n} N_A(\lambda^n T) \right) \left( \sup_{0 \leq t \leq T} (\tilde{W}^n(t))^2 \right) \Rightarrow 0, \tag{32}$$

as  $n \rightarrow \infty$  and hence the first part is proven.

Next, note that by Propositions 8.2 and 8.3 we have that  $\{\tilde{W}^n\}$  is tight and hence by Theorem 5.1 of [2] it is relatively compact. Let  $\{n_k\}$  be a subsequence along which  $\{\tilde{W}^n\}$  converges to some limit  $\tilde{W}$ . It then follows that we have since again as in the proof of Proposition 9.1,  $n^{-1} N_A(\lambda^n e) \Rightarrow$

$\lambda e$  as  $n \rightarrow \infty$ , we have by Theorem 3.9 of [2] that  $(\tilde{W}^{n_k}, n_k^{-1}N_A(\lambda^{n_k}e)) \Rightarrow (\tilde{W}, \lambda e)$  as  $n \rightarrow \infty$ . By the Skorohod Representation Theorem [2], there exists an alternate probability space on which we may assume that  $(\tilde{W}^{n_k}, n_k^{-1}N_A(\lambda^{n_k}e)) \rightarrow (\tilde{W}, \lambda e)$   $\mathbb{P}$ -a.s. It then follows by Lemma 8.3 of [6] and the fact that for each  $T \geq 0$  we have that  $\{\sup_{0 \leq t \leq T} |\tilde{W}^{n_k}(t)|\}$  is  $\mathbb{P}$ -a.s. bounded that

$$\begin{aligned} & \sup_{0 \leq t \leq T} \left| \frac{1}{n_k} \sum_{i=1}^{N_A(\lambda^{n_k}t)} \int_0^{\tilde{W}^{n_k}(\tau_i/\lambda^{n_k}-)} h(u)du - \lambda \int_0^t \int_0^{\tilde{W}^{n_k}(s)} h(u)duds \right| \\ &= \sup_{0 \leq t \leq T} \left| \int_0^t \int_0^{\tilde{W}^{n_k}(s-)} h(u)dud \left( \frac{1}{n_k} N_A(\lambda^{n_k}s) \right) - \lambda \int_0^t \int_0^{\tilde{W}^{n_k}(s)} h(u)duds \right| \\ &\rightarrow 0, \end{aligned}$$

as  $n \rightarrow \infty$ . This, then implies that

$$\frac{1}{n_k} \sum_{i=1}^{N_A(\lambda^{n_k}e)} \int_0^{\tilde{W}^{n_k}(\tau_i/\lambda^{n_k}-)} h(u)du - \lambda \int_0^e \int_0^{\tilde{W}^{n_k}(s)} h(u)duds \Rightarrow 0,$$

as  $n \rightarrow \infty$ . However, since the choice of  $\{n_k\}$  was arbitrary, the result follows.

Finally, note that by Proposition 8.3 we have that for each  $T \geq 0$ ,

$$\begin{aligned} & \sup_{0 \leq t \leq T} \left| \lambda \left( \int_0^t \int_0^{\tilde{W}^n(s)} h(u)duds - \int_0^t \int_0^{\tilde{Q}^{n,+}(s)} h(u)duds \right) \right| \\ &\leq \lambda \|h\|_\infty T \sup_{0 \leq t \leq T} |\tilde{W}^n(s) - \tilde{Q}^{n,+}(s)| \\ &\Rightarrow 0, \end{aligned}$$

as  $n \rightarrow \infty$  and hence the proof is complete.  $\square$

We may now provide the proof of Theorem 3.1.

*Proof of Theorem 3.1.* By (24) above and we have that

$$\begin{aligned} \tilde{Q}^n(t) &= \tilde{Q}_0^n + \tilde{N}_A(n^{-1}\lambda^n t) - \tilde{N}_D \left( \mu \int_0^t (\tilde{Q}^n(s) \wedge 1) ds \right) + \tilde{B}^n(t) + \tilde{\delta}^n(t) \\ &\quad + n^{1/2}(n^{-1}\lambda^n - \mu)t - \mu \int_0^t g(\tilde{Q}^n(s)) ds, \end{aligned}$$

where  $g(x) = \mu(x^- + \int_0^{x^+} h(u)du)$  for  $x \in \mathbb{R}$ . Since  $\|h\| < \infty$ , the function  $g$  is easily seen to be Lipschitz continuous and hence by Theorem 4.1 of [21] we have that  $\tilde{Q}^n = f(\tilde{Q}_0^n, \tilde{N}_A(n^{-1}\lambda^n e) - \tilde{N}_D(\mu \int_0^e (\tilde{Q}^n(s) \wedge 1) ds) + \tilde{B}^n(e) + \tilde{\delta}^n(e) + n^{1/2}(n^{-1}\lambda^n - \mu)e)$  where the function  $h : \mathbb{R} \times D \mapsto D$  is continuous with respect to both uniform and Skorohod  $J_1$  topology. In particular,  $f(x, y)$  is defined to be the unique solution to

$$z(t) = x + y(t) - \int_0^t g(z(s)) ds,$$

for  $t \geq 0$ . By Propositions 8.1 and 9.1 and Theorem 3.9 of [2] we have that  $(\tilde{Q}_0^n, \tilde{N}_A(n^{-1}\lambda^n e) - \tilde{N}_D(\mu \int_0^e (\tilde{Q}^n(s) \wedge 1) ds) + \tilde{B}^n(e) + \tilde{\delta}^n(e) + n^{1/2}(n^{-1}\lambda^n - \mu)e) \Rightarrow (\tilde{Q}_0, \sigma_A \tilde{B}_1(\mu e) + \tilde{B}_2(\mu e))$  as  $n \rightarrow \infty$ . The result now follows by the representation above and the Continuous Mapping Theorem [27].  $\square$

Finally, the proof of Proposition 3.2 is as follows.

*Proof of Proposition 3.2.* First note that the generator of  $\tilde{Q}$  is given for  $f \in C_b^2$  by

$$(\mathcal{A}f)(x) = \frac{\mu(1 + \sigma_A^2)}{2} f''(x) - \mu\beta f'(x) - 1\{x < 0\}\mu x f'(x) - 1\{x > 0\}\mu \left( \int_0^x h(u) du \right) f'(x)$$

It then follows that the stationary distribution  $\pi$  of  $\tilde{Q}$  must satisfy [9]

$$\int_{\mathbb{R}} \mathcal{A}f(x)\pi(dx) = 0,$$

for all  $f \in C_b^2(\mathbb{R})$ .

We now claim that  $\pi$  as given above satisfies this relationship. Integrating by parts that

$$\begin{aligned} & \int_{\mathbb{R}_+} \mathcal{A}f(x)\pi(dx) \\ &= \int_{\mathbb{R}_+} \left( \frac{\mu(1 + \sigma_A^2)}{2} f''(x) - \left( \mu\beta + \mu \left( \int_0^x h(u) du \right) \right) f'(x) \right) \exp \left( \frac{-2}{1 + \sigma_A^2} \left( \beta x + \int_0^x \int_0^u h(v) dv \right) \right) dx \\ &= \frac{\mu(1 + \sigma_A^2)}{2} f'(0) \\ & \quad + \int_{\mathbb{R}_+} \left( \mu f'(x) \left( \beta + \int_0^x h(u) du \right) - \left( \mu\beta + \mu \left( \int_0^x h(u) du \right) \right) f'(x) \right) \\ & \quad \times \exp \left( \frac{-2}{1 + \sigma_A^2} \left( \beta x + \int_0^x \int_0^u h(v) dv \right) \right) dx \\ &= \frac{\mu(1 + \sigma_A^2)}{2} f'(0). \end{aligned}$$

Similarly, for  $\mathbb{R}_-$  we obtain that

$$\begin{aligned} \int_{\mathbb{R}_-} \mathcal{A}f(x)\pi(dx) &= - \int_0^\infty \left( \frac{\mu(1 + \sigma_A^2)}{2} f''(x) - (\mu\beta + \mu x) f'(x) \right) \exp \left( \frac{-2}{1 + \sigma_A^2} (\beta x + x^2/2) \right) dx \\ &= - \frac{\mu(1 + \sigma_A^2)}{2} f'(0) \\ & \quad + \int_{\mathbb{R}_+} \left( \mu f'(x) (\beta + x) - (\mu\beta + \mu x) f'(x) \right) \exp \left( \frac{-2}{1 + \sigma_A^2} (\beta x + x^2/2) \right) dx \\ &= - \frac{\mu(1 + \sigma_A^2)}{2} f'(0). \end{aligned}$$

Thus,

$$\int_{\mathbb{R}} \mathcal{A}f(x)\pi(dx) = \int_{\mathbb{R}_-} \mathcal{A}f(x)\pi(dx) + \int_{\mathbb{R}_+} \mathcal{A}f(x)\pi(dx) = 0,$$

which completes the proof.  $\square$

## References

- [1] François Baccelli and Gérard Hebuterne. *Performance 81*, chapter On Queues with Impatient Customers, pages 159–179. North-Holland Publishing Company.
- [2] P. Billingsley. *Convergence of Probability Measures*. John Wiley and Sons, New York, 1999.
- [3] A. Brandt and M. Brandt. Asymptotic results and a markovian approximation for the  $M(n)/M(n)/s+GI$  system. *Queueing Systems*, 41:73–94, 2000.
- [4] A. Brandt and M. Brandt. On the  $m(n)/m(n)/s$  queue with impatient calls. *Performance Evaluation*, 35:1–18, 99.
- [5] G. Dai, S. He, and T. Tezcan. Many-server diffusion limits for  $G/Ph/n+GI$  queues. Technical report, Georgia Institute of Technology, 2009.
- [6] J. G. Dai and W. Dai. A heavy traffic limit theorem for a class of open queueing networks with finite buffers. *Queueing Systems: Theory and Applications*, 32:5–40, 1999.
- [7] J. G. Dai and Shuangchi He. Customer abandonment in many-server queues. Technical report, Georgia Institute of Technology, 2009.
- [8] C. H. Edwards and D. E. Penny. *Calculus with Analytic Geometry*. Prentice Hall, New Jersey, 1994.
- [9] S. Ethier and T. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, New York, 1986.
- [10] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, 1998.
- [11] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review and research prospects. *Manufacturing and Service Operations Management*, 5:79–141, 2003.
- [12] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.
- [13] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. 29(3):567–588, 1981.
- [14] R. Haugen and E. Skogan. Queueing systems with stochastic time out. *Communications, IEEE Transactions on*, 28(12):1984–1989, Dec 1980.
- [15] W. Kang and K. Ramanan. Fluid limits of many-server queues with reneging. *Preprint*, 2008.
- [16] A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: constraint satisfaction in call centers. Technical report, Technion, 2008.
- [17] M. Mandelbaum and P. Momcilovic. Queues with many servers and impatient customers. Technical report, Univeristy of Michigan–Ann Arbor, 2009.
- [18] Boxma O.J. and de Waal P.R. Multiserver queues with impatient customers. *ITC*, 14:743–756, 1994.

- [19] C. Palm. Methods of judging the annoyance caused by congestion. *Tele*, 4:189–208, 1953.
- [20] C. Palm. Research on telephone traffic carried by full availability groups. *Tele*, 1:107, 1957.
- [21] G. Pang, R. Talreja, and W. Whitt. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surv.*, 4:193–267, 2007.
- [22] J. E Reed and A. R. Ward. Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Mathematics of Operations Research*, 33:606–644, 2008.
- [23] M. I Reiman. The heavy-traffic limit theorem for sojourn times in Jackson networks. *Applied Probability - Computer Science, the Interface II*, R. L . Disney and T.J. Ott (editors):409–422, 1982.
- [24] Zeltyn S. and Mandelbaum A. Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. *QUESTA*, 51:361–402, 2005.
- [25] Amy R. Ward and Peter W. Glynn. A diffusion approximation for a markovian queue with reneging. *Queueing Syst. Theory Appl.*, 43(1/2):103–128, 2003.
- [26] Amy R. Ward and Peter W. Glynn. A diffusion approximation for a gi/gi/1 queue with balking or reneging. *Queueing Syst. Theory Appl.*, 50(4):371–400, 2005.
- [27] W. Whitt. *Stochastic-Process Limits*. Springer, New York, 2002.