Asymptotic Analysis of Queueing Systems with Reneging: A Survey of Results for FIFO, Single Class Models

Amy R. Ward

August 23, 2011

In this paper, we survey results for the $GI/GI/N + GI$ queueing model. Our focus is on finding situations in which simple performance measure approximations can be developed. To do this, we study the behavior of the $GI/GI/N + GI$ queue in the conventional heavy traffic and Halfin-Whitt limit regimes, and we also discuss the overloaded regime in which there is a single server as well as the overloaded many-server regime.

1 Introduction

Models of systems that involve customers queueing for service are called reneging models when there is the possibility that some customers may abandon the system before their service begins. Such models arise in a wide variety of application contexts, including call centers, communication networks, health care, and inventory systems. Reneging in the call center context occurs when a customer hangs up before reaching a service agent, and it is widely recognized that queueing models with reneging can help determine appropriate staffing levels; see the review papers of Aksin, Armony and Mehrotra [3] and Gans, Koole and Mandelbaum [32] and the references therein, and the recent work of Zhang, van Leeuwaarden, and Zwart [91]. Queueing models with reneging are also relevant for packet switched communications networks with time-critical traffic, in which each packet (customer) loses its value (reneges) if not transmitted (served) within a given time interval; see Bhattacharya and Ephremides [16]. This occurs, for example, when transmitting voice traffic, because delay must be smaller than 300 ms for normal conversation; see Panward, Towsley, and Wolf [66]. In the health care applications context, one major concern in emergency rooms is the presence of patients (customers) that leave without being seen (renege), and recent work of Green et al [39] has sought to use queueing models to reduce the percentage of patients that leave without being seen. Also in the health care applications context, Su and Zenios [77] model a kidney transplant waiting system as a queue with reneging, where reneging occurs because a customer that is waiting for a kidney may die. Finally, inventory systems that involve perishable goods also map to queueing models with reneging, as discussed in Nahmias [65]. This connection is made explicit in Kaspi and Perry [54] in the case of a blood bank.

1Marshall School of Business, University of Southern California, amyward@usc.edu
The commonality in all of the above applications is that analyzing a $GI/GI/N + GI$ queueing model can help to understand system performance. For example, it is of interest to know the steady-state distribution of the number of customers in the system, the probability that an arriving customer eventually reneges, and how the number of customers in the system evolves over time. However, exact analysis of $GI/GI/N + GI$ models is very difficult, and most exact analysis results require additional distributional assumptions (such as exponentiality) on the inter-arrival, service or reneging times. In fact, we are aware of no convenient, closed-form, exact expressions for performance measures for any $GI/GI/N + GI$ queueing model. This leads us to consider approximations.

We concentrate on $GI/GI/N + GI$ models in which the time at which each individual customer will renge is unknown. This is true in most any service environment, such as a call center or a health care setting, in which customers are people. In other applications settings, such as communication networks with time-critical traffic, or perishable commodity inventory systems, in which “customers” are packets or inventory units, it may or may not be true. Then, when there is a single class of customers, it is appropriate to schedule customers to be served in accordance with the FIFO service discipline. Otherwise, when individual customer reneging times are known, the service discipline that schedules customers in accordance with the time remaining before the customer will renge (more commonly known as the earliest deadline first, or EDF, discipline) is more appropriate (and has been shown to be optimal; see [66] and Section 5 in [56]).

Of course, in both call center and health care settings, as in many service application contexts, there are many customer classes. The classes are differentiated by, for example, their priority levels and/ or their impatience (how long each customer will wait before reneging). It is reasonable to assume FIFO service within each customer class; however, every time a server (that has the skills to serve more than one customer class) becomes free, he must choose which class to serve next. This leads to real-time scheduling control questions, with a partial list of papers that have studied such questions being: Ata and Rubino [5], Ghamami and Ward [35], and Ghosh and Weerasinghe [36] for models with a fixed number of servers, and Atar [7] [6], Atar, Giat, and Shimkin [9] [10], Atar, Mandelbaum, and Reiman [11], Gurvich and Whitt [42], Harrison and Zeevi [45], Perry and Whitt [68], and Tezcan and Dai [79] for models with a large number of servers. However, all of the aforementioned papers assume exponential reneging distributions, even though such an assumption may be unrealistic in practice. A first step towards solving such control questions without the assumption of exponential reneging times is understanding the behavior of the $GI/GI/N + GI$ model with a single customer class and FIFO service.

The focus of this paper is to survey diffusion approximations for $GI/GI/N + GI$ queueing models with a single customer class and FIFO service, and to understand how the diffusion approximations can be used to form simple expressions that estimate performance measures accurately. We concentrate on systems that are heavily loaded in the sense that the arrival rate to the system is very close to the system capacity, and also in which the mean time until a customer reneges is much larger than the mean inter-arrival time. Then, the diffusion processes that approximate the system are simple and very analytically tractable, except in
one case. That case occurs when the number of servers is large and service times are not exponential, and the reason is that the residual service times are a complicating factor that cannot be ignored.

The limit regimes we focus on to develop the diffusion approximations are the conventional heavy traffic (HT) limit regime and the Halfin-Whitt (HW) limit regime. In the conventional HT regime, the number of servers stays fixed while the arrival and service rates grow large, and, in the HW regime, the service rate stays fixed while the arrival rate and number of servers grows large. We also investigate an intermediate limit regime in which both the service rate and the number of servers grows large as the arrival rate grows large. Finally, we discuss the overloaded regime, and note that in that regime very simple performance measure approximations based on fluid limits can be extremely accurate.

It is of interest to have a single diffusion that in some sense unifies the diffusion approximations for the GI/GI/N + GI system that arise in the aforementioned limit regimes. We provide one step in this direction by presenting a diffusion whose limiting behavior is consistent with the behavior of the number-in-system process in a GI/M/N + GI queue in both the conventional HT and the HW limit regimes, and also in the intermediate limit regime.

The remainder of this paper is organized as follows. We begin in Section 2 by studying the M/M/N + M model. This is of interest in its own right; however, an even stronger motivation may be that understanding the M/M/N + M model provides intuition for how to approach the more general GI/GI/N + GI model. Section 3 we devote to developing results for the GI/GI/N + GI model. We show in Section 4 how a single diffusion can unify some of those results.

2 Exponential Models

The most basic queueing model that incorporates customer reneging is the M/M/N + M model with FIFO service. This model assumes that customers arrive to an infinite waiting room service facility according to a Poisson process, that their service times form an i.i.d. sequence of exponential random variables, and that each customer independently reneges if his service has not begun within an exponentially distributed amount of time. The four parameters that characterize this model are: the arrival rate \( \lambda \), the mean service time \( 1/\mu \), the number of servers \( N \), and the mean time a customer will wait before reneging \( 1/\gamma \).

We begin our analysis of the M/M/N + M model in Section 2.1 with some exact performance measure computations. Next, in Section 2.2, we establish the appropriate diffusion approximations for the number-in-system process in the conventional HT and the HW limit regimes, and also in an intermediate limit regime. Finally, in Section 2.3 we show how those diffusion approximations can be used to develop performance measure approximations.

2.1 Exact Analysis

The process \( Q = \{Q(t) : t \geq 0\} \) that tracks the number of customers in the system is a birth-death continuous time Markov chain with state space \( \mathbb{Z}^+ := \{0, 1, 2 \ldots\} \). The birth
rate is constant, and is $\lambda$. The death rate is state-dependent, and is $\mu \min(n, N) + \gamma[n - N]^+$ when there are $n$ customers in the system. This structure implies that the steady-state probabilities $\pi_n$ always exist and are

$$\pi_n = \begin{cases} \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n \pi_0, & n \in \{0, 1, \ldots, N\} \\ \frac{1}{N!} \left( \frac{\lambda}{\mu} \right)^N \left( \prod_{j=N+1}^n \frac{\lambda}{N \mu + (j-N)\gamma} \right) \pi_0, & n \in \{N + 1, N + 2, \ldots\} \end{cases},$$

where

$$\pi_0 = \left[ 1 + \sum_{n=1}^N \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n + \frac{1}{N!} \left( \frac{\lambda}{\mu} \right)^N \sum_{n=N+1}^{\infty} \prod_{j=N+1}^n \frac{\lambda}{N \mu + (j-N)\gamma} \right]^{-1}.$$

The steady-state probabilities in (2.1) can be expressed in terms of the gamma function $\Gamma(y) := \int_0^\infty t^{y-1} e^{-t} dt$, and the incomplete gamma function $I(x, k) = \int_x^{\infty} e^{-y} y^k dy / \Gamma(k+1)$, which was shown in Ancker and Gafarian [4] in the case that $N = 1$. To see this, first note that

$$\prod_{j=N+1}^n \frac{\lambda}{N \mu + (j-N)\gamma} = \frac{N \mu}{\lambda} \left( \frac{\lambda}{N \mu} \right) \prod_{j=N+1}^n \frac{\lambda/\gamma}{N \mu/\gamma + (j-N)}$$

$$= \frac{N \mu}{\lambda} \prod_{j=0}^{n-N} \frac{\lambda/\gamma}{N \mu/\gamma + j}$$

$$= \frac{N \mu}{\lambda} \left( \frac{\lambda}{\gamma} \right)^{n-N} \prod_{j=0}^{n-N} \frac{1}{N \mu/\gamma + j}.$$

Since it follows from the relation $\Gamma(z + 1) = z \Gamma(z)$ that

$$\prod_{j=0}^{n-N} \frac{1}{N \mu/\gamma + j} = \frac{\Gamma(N \mu/\gamma)}{\Gamma(N \mu/\gamma + n - N + 1)},$$

we find that

$$\prod_{j=N+1}^n \frac{\lambda}{N \mu + (j-N)\gamma} = \frac{N \mu}{\lambda} \left( \frac{\lambda}{\gamma} \right)^{n-N+1} \frac{\Gamma(N \mu/\gamma)}{\Gamma(N \mu/\gamma + n - N + 1)}.$$

The next step requires the relation in Eq. 12 in [4]

$$\sum_{k=0}^{\infty} \frac{x^k}{\Gamma(\alpha + k + 1)} = e^x x^{-\alpha} I \left( \frac{x}{\sqrt{\alpha}}, \alpha - 1 \right),$$

from which it follows that

$$\sum_{n=0}^{\infty} \frac{\left( \frac{\lambda}{\gamma} \right)^n}{\Gamma \left( \frac{N \mu}{\gamma} + n + 1 \right)} = \exp \left( \frac{\lambda}{\gamma} \right) \left( \frac{\lambda}{\gamma} \right)^{-\frac{N \mu}{\gamma}} I \left( \frac{\lambda/\gamma}{\sqrt{N \mu/\gamma}}, \frac{N \mu}{\gamma} - 1 \right).$$
Then, since 
\[
\pi_0 = \left[ 1 + \sum_{n=1}^{N-1} \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n + \frac{1}{N!} \left( \frac{\lambda}{\mu} \right)^N \left( \frac{N\mu}{\lambda} \right) \Gamma \left( \frac{N\mu}{\gamma} \right) \Gamma \left( \frac{\lambda}{\gamma} \right) \sum_{n=0}^{\infty} \frac{\left( \frac{\lambda}{\gamma} \right)^n}{\Gamma \left( \frac{N\mu}{\gamma} + n + 1 \right)} \right]^{-1},
\]
equation (2.1) is equivalently written as
\[
\pi_n = \begin{cases} 
\frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n \pi_0, & n \in \{0, 1, \ldots, N\} \\
\frac{1}{N!} \left( \frac{\lambda}{\mu} \right)^N \left( \frac{N\mu}{\lambda} \right) \Gamma \left( \frac{N\mu}{\gamma} \right) \Gamma \left( \frac{\lambda}{\gamma} \right) \sum_{n=0}^{\infty} \frac{\left( \frac{\lambda}{\gamma} \right)^n}{\Gamma \left( \frac{N\mu}{\gamma} + n + 1 \right)} \pi_0, & n \in \{N + 1, N + 2, \ldots\}
\end{cases}
\]
where
\[
\pi_0 = \left[ 1 + \sum_{n=1}^{N-1} \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n + \frac{1}{N!} \left( \frac{\lambda}{\mu} \right)^N \left( \frac{N\mu}{\lambda} \right) \Gamma \left( \frac{N\mu}{\gamma} \right) \Gamma \left( \frac{\lambda}{\gamma} \right) \sum_{n=0}^{\infty} \frac{\left( \frac{\lambda}{\gamma} \right)^n}{\Gamma \left( \frac{N\mu}{\gamma} + n + 1 \right)} \right]^{-1}.
\]
Equation (2.2) is useful for computing the steady-state probabilities, because the gamma and incomplete gamma functions are built-in functions in many technical computing software packages, such as Mathematica.

This is a convenient place to briefly detour, and remark on a notational convention. Here and throughout this paper, for any stochastic process \( \{X(t) : t \geq 0\} \), we let \( X(\infty) \) denote the random variable that has the steady-state distribution associated with \( X \) (assuming that there exists a unique steady-state distribution). So, for example, \( Q(\infty) \) has the steady-state distribution associated with \( Q \), that is defined by \( \pi_n \).

Most any steady-state performance measure for the \( M/M/N+M \) model can be expressed in terms of the steady-state probabilities in (2.2). For example, the expected number of customers in the system is
\[
E[Q(\infty)] = \sum_{q=0}^{\infty} q \pi_q.
\]
The steady-state abandonment probability \( P_a \) (that is, the probability that an arriving customer reneges and is not served) can be computed through the relation
\[
\gamma E[(Q(\infty) - N)^+] = \lambda P_a,
\]
which states that the steady-state rate at which customers reneging must equal the steady state rate at which customers that eventually abandon enter the system. Then,
\[
P_a = \frac{\gamma}{\lambda} \sum_{q=N+1}^{\infty} (q - N) \pi_q.
\]
When \( \lambda/(N\mu) < 1 \), an alternative formula can be found in Theorem 1 in Boots and Tijms [19], and is
\[
P_a = \frac{\left( 1 - \frac{\lambda}{N\mu} \right) P(W_{M/M/N}(\infty) > \tau)}{1 - \frac{\lambda}{N\mu} P(W_{M/M/N}(\infty) > \tau)},
\]
5
where $\tau$ has the distribution of an exponential random variable with mean $1/\gamma$ and $W_{M/M/N}(\infty)$ has the distribution of the steady-state waiting time in an $M/M/N$ queue.

We may also compute the steady-state expected time an arrival spends in the system conditional on that arrival receiving service, following Whitt [85]. First note that if $S$ is the event that an arbitrary arrival receives service, then

$$P(S) = \sum_{q=0}^{\infty} P(S|Q(\infty) = q)\pi_q.$$  

Since when there are $i$ waiting customers in the system ahead of a given tagged customer, the probability the next event is that tagged customer reneging is $\gamma/(N\mu + (i + 1)\gamma)$, it follows that

$$P(S|Q(\infty) = q) = \prod_{i=0}^{q-N} \left(1 - \frac{\gamma}{N\mu + (i + 1)\gamma}\right) = \frac{N\mu}{N\mu + (q - N + 1)\gamma}.$$  

Also, the steady-state expected time in the system, $E[W(\infty)]$, for an infinitely patient arrival is

$$E[W(\infty)] = \frac{1}{\mu} \sum_{q=0}^{\infty} \pi_q + \sum_{q=N}^{\infty} \left(\sum_{i=1}^{q-N+1} \frac{1}{N\mu + i\gamma} + \frac{1}{\mu}\right)\pi_q.$$  

In summary,

$$E[W(\infty)|S] = \frac{E[W(\infty)]}{P(S)}.$$  

For more computations (specifically, for the mean and variance of both the conditional time to complete service given that service is completed and the conditional time to renege given that the customer reneges, and also for the Laplace transforms of the full conditional distributions), see Section 3 in [85].

The previous paragraph motivates having an analytically tractable process that has similar transient and steady-state distributional behavior as $Q$. This can be done by finding limiting regimes in which there arise diffusion process approximations for $Q$. It turns out that the snapshot principle (a Little’s law that holds at all times $t > 0$; see Reiman [74] for the result in the case of a Jackson network) holds in those limit regimes, so that knowledge of $Q$ translates to knowledge about waiting; see Remark 3.2. Hence it is enough to focus on the process $Q$.

---

2 Although [85] assumes an $M/M/N+M$ queue with a finite buffer size, the computations can be modified in a straightforward manner to accommodate an infinite buffer.
2.2 Limit Theorems for the $M/M/N + M$ Queue

We re-state the limit theorems for the $M/M/N + M$ queue in both the conventional HT regime and in the HW many-server limit regime. When $\lambda$ becomes large, these two limit regimes arise by letting $\mu$ or $N$ also become large.

**Definition 2.1** The Conventional Heavy Traffic (HT) Limit Regime.
For a given $\beta \in \mathbb{R}$, define $\mu^\lambda := (\lambda + \beta \sqrt{\lambda})/N$, and assume that $N$ is fixed and independent of $\lambda$.

**Definition 2.2** The Halfin Whitt (HW) Limit Regime.
For a given $\beta \in \mathbb{R}$, define $N^\lambda := (\lambda + \beta \sqrt{\lambda})/\mu$, and assume that $\mu$ is fixed and independent of $\lambda$.

In both the conventional HT and HW limit regimes\(^3\), the system loading factor $\rho := \lambda/N\mu$ approaches 1, since it follows from Definition 2.1 that

$$\rho^\lambda = \frac{\lambda}{N\mu^\lambda} = \frac{\lambda}{\lambda + \beta \sqrt{\lambda}} \to 1, \text{ as } \lambda \to \infty,$$

and it follows from Definition 2.2 that

$$\rho^\lambda = \frac{\lambda}{N^\lambda \mu} = \frac{\lambda}{\lambda + \beta \sqrt{\lambda}} \to 1 \text{ as } \lambda \to \infty.$$

The key conceptual difference between these two regimes is that in conventional HT, an arriving customer is almost certain to be delayed, but in the HW regime, the steady-state probability that an arriving customer is delayed remains strictly between 0 and 1. This is discussed in detail in the paper by Halfin and Whitt [43] in the case of the $M/M/N$ queue, and is a primary motivation for their development of the HW limit regime. Garnett, Mandelbaum, and Reiman [33] show that it continues to hold that the steady-state probability that an arriving customer is delayed is strictly between 0 and 1 in the case of an $M/M/N + M$ queue in the Halfin-Whitt regime.

There are also a range of intermediate limit regimes that arise by letting both $\mu$ and $N$ become large as $\lambda$ becomes large.

**Definition 2.3** An Intermediate Limit Regime.
For a given $\beta \in \mathbb{R}$, define $\mu^\lambda = \mu \left(\lambda^{1-\alpha} + \lambda^{1/2-\alpha} \beta \right)$ and $N^\lambda = \lambda^\alpha/\mu$ for $\alpha \in [0, 1]$.

\(^3\)The conventional HT limit regime is often defined using a sequence of systems, indexed by $n$, and letting $\lambda$ and $\mu$ be order 1 quantities. Also, the HW limit regime is often defined by letting $N$ become large, and having $\lambda$ depend on $N$. The presentation we have chosen (1) means that we do not need to time scale any system processes to obtain weak convergence results (whereas, in the aforementioned more common conventional heavy traffic definition, as in, for example, Iglehart and Whitt [46], a time scaling is necessary), and (2) both the conventional HT and HW limit regimes are defined by letting the *same* parameter ($\lambda$) become large.
We also describe the limiting behavior of the $M/M/N + M$ queue in the intermediate limit regime of Definition 2.3. Note that if $\alpha = 1$ this is like the HW limit regime with $\beta = 0$, and if $\alpha = 0$ it is like the conventional HT limit regime. Again,

$$\rho^\lambda = \frac{\lambda}{N^\lambda \mu^\lambda} \to 1,$$

as $\lambda \to \infty$.

To motivate the limit theorems, it is useful to first calculate the infinitesimal mean and variance of the process $Q$. This is because in a limit regime in which a diffusion approximation for $Q$ arises, these infinitesimal parameters, when appropriately scaled, will converge to the infinitesimal parameters of that diffusion. The process $Q$ has infinitesimal mean

$$m_Q(n) := \lim_{h \downarrow 0} \frac{1}{h} E \left[ Q^\lambda(h) - Q^\lambda(0) \bigg| Q^\lambda(0) = n \right]$$

and infinitesimal variance

$$v_Q(n) := \lim_{h \downarrow 0} \frac{1}{h} E \left[ (Q^\lambda(h) - Q^\lambda(0))^2 \bigg| Q^\lambda(0) = n \right]$$

2.2.1 The $M/M/N + M$ Queue in the Conventional HT Regime

We calculate the infinitesimal mean and variance of the process $Q^\lambda/\sqrt{\lambda}$, and then let $\lambda \to \infty$ in the conventional HT regime in Definition 2.1. Recalling that $\mu^\lambda = (\lambda + \beta \sqrt{\lambda})/N$ for $N$ fixed, note that when $\lambda > (N/x)^2$,

$$\frac{1}{\sqrt{\lambda}} m_Q(\sqrt{\lambda} x) = \frac{1}{\sqrt{\lambda}} \left( \lambda - N \mu^\lambda + \mu^\lambda \left[ N - \sqrt{\lambda} x \right]^+ - \gamma \left[ \sqrt{\lambda} x - N \right]^+ \right) = -(\beta + \gamma x),$$

for $x > 0$. Hence

$$\lim_{h \downarrow 0} \frac{1}{h} E \left[ \frac{Q^\lambda(h)}{\sqrt{\lambda}} - \frac{Q^\lambda(0)}{\sqrt{\lambda}} \bigg| \frac{Q^\lambda(0)}{\sqrt{\lambda}} = x \right] \to -\beta - \gamma x,$$

as $\lambda \to \infty$. Similarly,

$$\lim_{h \downarrow 0} \frac{1}{h} E \left[ \left( \frac{Q^\lambda(h)}{\sqrt{\lambda}} - \frac{Q^\lambda(0)}{\sqrt{\lambda}} \right)^2 \bigg| \frac{Q^\lambda(0)}{\sqrt{\lambda}} = x \right] \to \frac{1}{\lambda} v_Q(\sqrt{\lambda} x) \to 2,$$

as $\lambda \to \infty$, for $x > 0$. This calculation suggests that $Q^\lambda/\sqrt{\lambda}$ converges to the reflected Ornstein-Uhlenbeck diffusion $X_{ROU} = \{X_{ROU}(t) : t \geq 0\}$ that has infinitesimal drift

$$m_{ROU}(x) := -\beta - \gamma x \text{ for } x > 0,$$
infinitesimal variance $\sigma^2 = 2$, and is the strong solution to the stochastic equation

$$X_{\text{ROU}}(t) = X_{\text{ROU}}(0) + \int_0^t m_{\text{ROU}}(X(s)) \, ds + \sigma B(t) + L_{\text{ROU}}(t),$$  \hspace{1cm} (2.4)

where $X_{\text{ROU}}(0) \geq 0$, $B$ is a standard Brownian motion, and $L_{\text{ROU}}$ is the minimal non-decreasing process under which $X_{\text{ROU}}(t) \geq 0$ for all $t \geq 0$. The following result, which is a restatement of Theorem 1 part 1 in Ward and Glynn [82], but converted to our setting, establishes the above intuition rigorously. The result is a weak convergence result in $D[0, \infty)$, as are all the weak convergence results in this paper, and we refer the reader to Billingsley [17] or Whitt [86] for more on this convergence concept.

**Theorem 2.1 (Conventional Heavy Traffic Regime)**

Let $Q^{\lambda}$ denote the number-in-system process for the $M/M/N + M$ model having parameters $\lambda, \mu^\lambda$, and $N$, given as in the conventional heavy traffic limit regime in Definition 2.1, and $\gamma$ fixed. Then,

$$\frac{Q^{\lambda}}{\sqrt{\lambda}} \Rightarrow X_{\text{ROU}}, \text{ as } \lambda \to \infty, \text{ in } D[0, \infty).$$

We note that Theorem 2.1 can also be proved using Theorems 4.1 and 4.2 in Mandelbaum and Pats [62]. The paper [62] provides a unifying framework for finding fluid and diffusion limits for queues with Poisson arrival processes and exponential service times in which the birth and death rates in the associated CTMC may depend on the state. (Section 5.8 in [62] provides the closest illustrative example, which considers the queue with processor-shared service and reneging that was analyzed in Coffman et al [50]. For a more recent analysis of a processor-sharing queue with reneging, see Gromoll et al [40].)

### 2.2.2 The $M/M/N + M$ Queue in the HW Regime

We also calculate the infinitesimal mean and variance of the process $(Q^{\lambda} - N^\lambda) / \sqrt{\lambda}$, and then let $\lambda \to \infty$ in the HW limit regime in Definition 2.2. Recalling that $N^\lambda = (\lambda + \beta \sqrt{\lambda}) / \mu$ for $\mu$ fixed, note that

$$\frac{1}{\sqrt{\lambda}} m_Q \left( \sqrt{\lambda} x + N^\lambda \right) = \frac{1}{\sqrt{\lambda}} \left( \lambda - N^\lambda \mu + \mu \left[ -\sqrt{\lambda} x \right]^+ - \gamma \left[ \sqrt{\lambda} x \right]^+ \right)$$

$$= -\beta + \mu [-x]^+ - \gamma [x]^+, \hspace{1cm} \text{for } x \in \mathbb{R}.$$  

Hence

$$\lim_{h \downarrow 0} \frac{1}{h} E \left[ \frac{Q^\lambda(h) - N^\lambda}{\sqrt{\lambda}} - \frac{Q^\lambda(0) - N^\lambda}{\sqrt{\lambda}} \right] = \frac{Q^\lambda(0) - N^\lambda}{\sqrt{\lambda}} = x \to m_{\text{HW}}(x),$$

as $\lambda \to \infty$, where

$$m_{\text{HW}}(x) := \begin{cases} 
-\beta - \gamma x, & \text{if } x > 0 \\
-\beta - \mu x, & \text{if } x \leq 0 
\end{cases}.$$
Similarly,

\[
\lim_{h \to 0} \frac{1}{h} E \left[ \left( \frac{Q^\lambda(h) - N^\lambda}{\sqrt{\lambda}} - \frac{Q^\lambda(0) - N^\lambda}{\sqrt{\lambda}} \right)^2 \mid \frac{Q^\lambda(0) - N^\lambda}{\sqrt{\lambda}} = x \right] = \frac{1}{\lambda} v_Q \left( \sqrt{\lambda} x + N^\lambda \right) \to 2,
\]

as \( \lambda \to \infty \), for \( x \in \mathbb{R} \). This calculation suggests that \( (Q^\lambda - N^\lambda)/\sqrt{\lambda} \) converges to the diffusion \( X_{HW} = \{X_{HW}(t) : t \geq 0\} \) that has infinitesimal drift \( m_{HW} \) and infinitesimal variance \( \sigma^2 = 2 \), and is the strong solution to the stochastic equation

\[
X_{HW}(t) = X_{HW}(0) + \int_0^t m_{HW}(X(s)) \, ds + \sigma B(t), \tag{2.5}
\]

where \( X_{HW}(0) \in \mathbb{R} \), and \( B \) is a standard Brownian motion. This intuition is proved in Theorem 2 in [33], which we re-state here.

**Theorem 2.2 (Halfin-Whitt Regime)**

Let \( Q^\lambda \) denote the number-in-system process for the \( M/M/N + M \) model having parameters \( \lambda, \mu \) fixed, and \( N^\lambda \), given as in the HW limit regime in Definition 2.2, and \( \gamma \) fixed. Then,

\[
\frac{Q^\lambda - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{HW}, \text{ as } \lambda \to \infty, \text{ in } D[0, \infty).
\]

We observe that the infinitesimal drift of \( X_{ROU} \) matches the infinitesimal drift of \( X_{HW} \) in the positive part of the state space. In other words, when there are customers waiting, the transition dynamics of the two limit processes are exactly the same.

**Remark 2.1** When \( \gamma = 0 \), Theorem 2.1 is consistent with results on the conventional single-server queue in heavy traffic (see, for example, the books by Chen and Yao [25] or Whitt [86] for such results and much more), and Theorem 2.2 becomes a re-statement of Theorem 2 in [43]. Glynn [37] provides a nice, brief overview of diffusion approximation theory, as it applies to queues in which there is no reneging.

**Remark 2.2** We have assumed that customers in service do not renege. This makes sense in most service applications, because customers are aware of whether or not they are being served. However, sometimes it may be appropriate to assume that customers may renege while in service. (For example, such an assumption may be reasonable in the communications network example mentioned in the first paragraph of the Introduction.) Then, the death rates in the relevant continuous time Markov chain become \( \mu \min(n, N) + \gamma n \) when there are \( n \) customers in the system. The change to the death rates has no effect in the conventional HT regime, and Theorem 2.1 holds as stated. However, this is not the case in the HW limit regime (as can be seen by calculating the limiting infinitesimal mean and variance of the appropriately scaled process). In particular, Theorem 2.2 is no longer valid.
2.2.3 The $M/M/N + M$ Queue in an Intermediate Limit Regime

Finally, we calculate the infinitesimal mean and variance of the process $(Q^\lambda - N^\lambda)/\sqrt{\lambda}$, and then let $\lambda \to \infty$ in the intermediate limit regime in Definition 2.3. Recalling that $N^\lambda = \lambda^\alpha/\mu$ and $\mu^\lambda = \mu \left( \lambda^{1-\alpha} + \lambda^{1/2-\alpha}\beta \right)$ for $\alpha \in [0, 1]$, note that

$$
\lim_{h \downarrow 0} \frac{1}{h} E \left[ \frac{Q^\lambda(h) - N^\lambda}{\sqrt{\lambda}} - \frac{Q^\lambda(0) - N^\lambda}{\sqrt{\lambda}} \right] = \frac{1}{\sqrt{\lambda}} m_Q \left( \sqrt{\lambda}x + N^\lambda \right),
$$

as $\lambda \to \infty$, and

$$
\lim_{h \downarrow 0} \frac{1}{h} E \left[ \left( \frac{Q^\lambda(h) - N^\lambda}{\sqrt{\lambda}} - \frac{Q^\lambda(0) - N^\lambda}{\sqrt{\lambda}} \right)^2 \right] = \frac{1}{\sqrt{\lambda}} \mu_Q \left( \sqrt{\lambda}x + N^\lambda \right),
$$

as $\lambda \to \infty$. This calculation suggests that $(Q^\lambda - N^\lambda)/\sqrt{\lambda}$ converges to the reflected diffusion $X_{ROU}$ defined in (2.4), and this is proved in Theorem 2.1 in Atar [8] in a more general setting in which the service rates may be different for each server.

**Theorem 2.3** Let $Q^\lambda$ denote the number-in-system process for the $M/M/N + M$ model having parameters $\lambda, \mu^\lambda$, and $N^\lambda$, given as in the intermediate limit regime in Definition 2.3, and $\gamma$ fixed. Then,

$$
\frac{Q^\lambda - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{ROU}, \text{ as } \lambda \to \infty, \text{ in } D[0, \infty).
$$

2.3 Performance Measure Approximation

Theorems 2.1-2.3 validate approximating the number-in-system process in the $M/M/N + M$ model as

$$
Q(\cdot) \overset{D}{\approx} \sqrt{\lambda} X_{ROU}(\cdot), \quad (2.6)
$$

when $\lambda$ and $\mu$ are large, $N$ is small, and $\lambda$ is large compared to $\gamma$, as

$$
Q(\cdot) \overset{D}{\approx} \sqrt{\lambda} X_{HW}(\cdot) + N, \quad (2.7)
$$

as $\lambda \to \infty$. This calculation suggests that $(Q^\lambda - N^\lambda)/\sqrt{\lambda}$ converges to the reflected diffusion $X_{ROU}$ defined in (2.4), and this is proved in Theorem 2.1 in Atar [8] in a more general setting in which the service rates may be different for each server.

**Theorem 2.3** Let $Q^\lambda$ denote the number-in-system process for the $M/M/N + M$ model having parameters $\lambda, \mu^\lambda$, and $N^\lambda$, given as in the intermediate limit regime in Definition 2.3, and $\gamma$ fixed. Then,

$$
\frac{Q^\lambda - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{ROU}, \text{ as } \lambda \to \infty, \text{ in } D[0, \infty).
$$

2.3 Performance Measure Approximation

Theorems 2.1-2.3 validate approximating the number-in-system process in the $M/M/N + M$ model as

$$
Q(\cdot) \overset{D}{\approx} \sqrt{\lambda} X_{ROU}(\cdot), \quad (2.6)
$$

when $\lambda$ and $\mu$ are large, $N$ is small, and $\lambda$ is large compared to $\gamma$, as

$$
Q(\cdot) \overset{D}{\approx} \sqrt{\lambda} X_{HW}(\cdot) + N, \quad (2.7)
$$

as $\lambda \to \infty$. This calculation suggests that $(Q^\lambda - N^\lambda)/\sqrt{\lambda}$ converges to the reflected diffusion $X_{ROU}$ defined in (2.4), and this is proved in Theorem 2.1 in Atar [8] in a more general setting in which the service rates may be different for each server.
when $\lambda$ and $N$ are large, $\mu$ may not be large, and $\lambda$ is large compared to $\gamma$, and as

$$Q(\cdot) \xrightarrow{D} \sqrt{\lambda}X_{ROU}(\cdot) + N,$$  \hspace{2cm} (2.8)

when $\lambda$ is large, $\mu$ and $N$ are both medium, and $\lambda$ is large compared to $\gamma$. Then, (2.6)-(2.8) immediately imply that $X_{ROU}$ and $X_{HW}$ can be used to develop transient performance measure approximations for $Q$. For example, for a given $t > 0$ and $q > 0$, $E[Q(t)|Q(0) = q]$ will be close to $\sqrt{\lambda}E[X_{ROU}(t)|X_{ROU}(0) = q/\sqrt{\lambda}]$ when use of (2.6) is appropriate, and to $\sqrt{\lambda}E[X_{HW}(t)|X_{HW}(0) = (q - N)/\sqrt{\lambda}] + N$ when use of (2.7) is appropriate, and to $\sqrt{\lambda}E[X_{ROU}(t)|X_{ROU}(0) = (q - N)/\sqrt{\lambda}]$ when use of (2.8) is appropriate.

It is natural to expect that (2.6)-(2.8) also provide approximations for the steady-state distribution of $Q$, so that

$$Q(\infty) \xrightarrow{D} \sqrt{\lambda}X_{ROU}(\infty),$$  \hspace{2cm} (2.9)

when use of (2.6) is appropriate, and

$$Q(\infty) \xrightarrow{D} \sqrt{\lambda}X_{HW}(\infty) + N,$$  \hspace{2cm} (2.10)

when use of (2.7) is appropriate, and

$$Q(\infty) \xrightarrow{D} \sqrt{\lambda}X_{ROU}(\infty) + N,$$  \hspace{2cm} (2.11)

when use of (2.8) is appropriate. However, Theorems 2.1-2.3 do not rigorously justify the approximations in (2.9)-(2.11). In particular, it is still required to show that the sequence of steady-state distributions associated with $Q$, when appropriately scaled, converge to the steady-state distribution associated with $X_{ROU}$ and $X_{HW}$; that is, that

$$\frac{Q^\lambda(\infty)}{\sqrt{\lambda}} \Rightarrow X_{ROU}(\infty), \text{ as } \lambda \to \infty,$$

in the conventional HT limit regime in Definition 2.1, and

$$\frac{Q^\lambda(\infty) - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{HW}(\infty), \text{ as } \lambda \to \infty,$$

in the HW limit regime in Definition 2.2, and

$$\frac{Q^\lambda(\infty) - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{ROU}(\infty), \text{ as } \lambda \to \infty,$$

in the intermediate limit regime in Definition 2.3. This has been done in the conventional HT limit regime (see Proposition 1 in [82]) and also in the HW limit regime (see Fleming et al [31]). This remains to be established in the intermediate limit regime.

In summary, the development of simple performance measure approximations for the $M/M/N + M$ system depends on $X_{ROU}$ and $X_{HW}$ being analytically tractable processes.
First, their steady-state distributions can be expressed in terms of the standard normal pdf \( \phi \) and cdf \( \Phi \). Specifically, the steady-state density \( f_{ROU} \) of \( X_{ROU} \) is

\[
f_{ROU}(x) = \frac{\sqrt{2\gamma}}{\sigma} \phi \left( \frac{\sqrt{2\gamma}}{\sigma} \left(x + \frac{\beta}{\gamma}\right) \right) \left(1 - \Phi \left( \frac{\sqrt{2\beta} \sigma}{\gamma} \right) \right), \quad \text{for } x \geq 0,
\]

and the steady-state density \( f_{HW} \) of \( X_{HW} \) is

\[
f_{HW}(x) = \begin{cases} 
\frac{\sqrt{2\gamma}}{\sigma} \phi \left( \frac{\sqrt{2\gamma}}{\sigma} \left(x + \frac{\beta}{\gamma}\right) \right) p, & \text{if } x \leq 0, \\
\frac{\sqrt{2\gamma}}{\sigma} \phi \left( \frac{\sqrt{2\gamma}}{\sigma} \left(x + \frac{\beta}{\gamma}\right) \right) (1 - p), & \text{if } x > 0,
\end{cases}
\]

where

\[
p = \left[ 1 + \sqrt{\frac{\mu}{\gamma}} \phi \left( \frac{\sqrt{2\beta} \sigma}{\mu \sigma^2} \right) \left(1 - \Phi \left( \frac{\sqrt{2\beta} \sigma}{\gamma} \right) \right) \right]^{-1}
\]

see [24]. Then, it can be argued that (2.9)-(2.11) lead to more convenient expressions for steady-state performance measures, such as the mean steady-state queue-length, than those based on the exact steady-state probabilities in (2.2). Also, the relationship in (2.3) leads to simple expressions that approximate the abandonment probability, as follows

\[
\begin{align*}
P_a &= \frac{\gamma}{\lambda} E \left[ (Q(\infty) - N)^+ \right] \\
&\approx \begin{cases} 
\frac{\gamma}{\lambda} E \left[ X_{ROU}(\infty) \right] & \text{when use of (2.6) or (2.8) is appropriate} \\
\frac{\gamma}{\lambda} E \left[ [X_{HW}(\infty)]^+ \right] & \text{when use of (2.7) is appropriate}.
\end{cases}
\end{align*}
\]

The transient distribution of \( X_{ROU} \) is available via its Sturm-Liouville spectral expansion, as shown in Linetsky [58]. Also, there is an exact expression for the transient distribution of \( X_{ROU} \) in terms of the hitting time density of an unreflected O-U process in Cox and Rosler [26] (see Theorem 5 and the example at the top of page 154; however, the hitting time density of an unreflected O-U process cannot always be represented using convenient closed-form expressions. When \( \gamma \) is small, we can compute an approximation for the transition density and any transient moments of \( X_{ROU} \) by viewing it as a perturbation of an RBM process; see Theorem 1 in Ward and Glynn [83], and the computations immediately following. This approach yields convenient analytic expressions because there is a closed form expression for the RBM transition density in Harrison [44], and very good approximations for its transient moments can be found in Abate and Whitt [1] [2]. The Laplace transform of the transient distribution of \( X_{HW} \) is given in Knessl and van Leeuwaarden (2008) [81] when \( \gamma > 0 \) and Knessl and van Leeuwaarden (2010) [80] when \( \gamma = 0 \).

Regarding delay, the time a customer must wait before being served, the astute reader has observed that since queue size is of order \( \sqrt{\lambda} \) in both the conventional HT and HW
regimes, Little’s law suggests that expected steady-state delay in the system with arrival rate $\lambda$ is of order $1/\sqrt{\lambda}$. Then, in the conventional HT regime, delay is much larger than service time (since service times are of order $1/\lambda$) and in the HW regime, delay is much smaller than service time (since service times are of order 1). It is of interest to search for a limit regime in which delay and service times are of the same order. This occurs in the intermediate limit regime in Definition 2.3 when $\alpha = 1/2$, as established in Theorem 2.2 in [8]. That regime is called the non-degenerate slowdown regime in [8], because it is only when $\alpha = 1/2$ that the slowdown (defined as the ratio between the sum of the delay and the service time) approaches a finite value that is strictly greater than 1. In the conventional HT regime, the slowdown approaches $\infty$ as $\lambda \to \infty$, and in the HW regime, the slowdown approaches 1 as $\lambda \to \infty$. We remark that the non-degenerate slowdown regime was earlier considered in Mandelbaum [60] and Whitt [87] in the case of a $M/M/N$ model (without reneging) and also in Gurvich [41] in the case of a $M/M/N$ model that has many customer classes. The key refinement in [8] is the weak convergence of the joint law of delay and service time.

The importance of the approximations in (2.6)-(2.11) is not fully realized in the $M/M/N + M$ setting, because exact analysis computations for steady-state performance measures is possible. Their full power lies in the fact that the limit diffusion processes that arise in Theorems 2.1 and 2.2 also arise in non-exponential reneging models, so that the performance measure approximations developed in this subsection remain relevant in much more general models, in which exact analysis computations for both steady-state and transient performance measures appears, in general, prohibitively difficult.

3 Non-Exponential Models

The ideal is to have simple diffusion processes that arise as limits in much more general reneging models. In such models, the only possible way to understand system performance may be through approximations. We devote this section to developing performance measure approximations for much more general reneging models. We consider only the conventional HT and HW limit regimes because these are the regimes in which current work is focused.

We first construct a $GI/GI/N + GI$ model in Section 3.1, and then, in Section 3.2, show when the $X_{ROU}$ and $X_{HW}$ processes in (2.4) and (2.5) arise as limits. We next provide limit theorems, in Section 3.3, that serve as a basis for constructing performance measure approximations that can be more accurate than those based on the $X_{ROU}$ and $X_{HW}$ processes. Finally, in Section 3.4, we show that very simple performance measure approximations can arise in an overloaded regime, in which the arrival rate remains strictly above the total available service rate, or service capacity (instead of being equal to the service capacity, as is the case in the conventional HT and HW limit regimes).
3.1 The \( GI/GI/N + GI \) Model

We construct our \( GI/GI/N + GI \) queueing model from three independent sequences of i.i.d. random variables, \( \{u_i, i \geq 1\} \), \( \{v_i, i \geq 1\} \), and \( \{w_i, i \geq 1\} \), all defined on a common probability space \( (\Sigma, \mathcal{F}, P) \). We assume \( E[u_1] = E[v_1] = 1 \), and \( \text{var}(u_1) < \infty, \text{var}(v_1) < \infty \). For a given arrival rate \( \lambda \), the \( \text{ith} \) customer arrives to the queue at time \( \sum_{j=1}^{\text{i}} u_j/\lambda \), has service time \( v_i/\mu \), and reneges if his service does not begin within \( w_i \) time units. Customers are served in the order of their arrival (FIFO). We assume that \( F \) is proper (that is, \( F(x) \to 1 \) as \( x \to \infty \)), so that the system is stable for any values of \( \lambda \) and \( \mu \). The parameters that will be required to state our limit theorems for the \( GI/GI/N + GI \) model are:

- the arrival rate \( \lambda \);
- the squared coefficient of variation of the inter-arrival times \( c_a^2 := \text{var}(u_1/\lambda)/(1/\lambda)^2 = \text{var}(u_1) \);
- the service rate \( \mu \);
- the squared coefficient of variation of the service times \( c_s^2 := \text{var}(v_1/\mu)/(1/\mu)^2 = \text{var}(v_1) \);
- the number of servers \( N \);
- the reneging distribution \( F \).

The exact analysis of the \( GI/GI/N + GI \) model is very difficult. There are some exact results available, and those often require additional distributional assumptions (see, for example, Baccelli, Boyer, and Hebuterne [12], Bae, Kim, and Lee [13], Barrer [14], Boots and Tijms [19], Boxma, Perry, and Stadje [20], Boxma, Perry, Stadje, and Zacks [21], Brandt and Brandt [22] [23], Finch [30], Gavish and Schweitzer [34], Gnedenko and Kovalenko [38], Jurkevic [51], Movaghar [64], Perry and Asmussen [67], Rao [69], and Stanford [75]). However, in general, obtaining convenient closed form expressions for either steady-state or transient performance measures is not possible. Therefore, it is of interest to determine when the diffusions \( X_{ROU} \) and \( X_{HW} \) that arise as approximations for the \( M/M/N + M \) model can be used to approximate non-exponential reneging models, and if there are other analytically tractable approximations for such models.

3.2 Limit Theorems in which \( X_{ROU} \) and \( X_{HW} \) Arise

We begin with a result that establishes that \( X_{ROU} \) can be an appropriate approximation for a \( GI/GI/N + GI \) model. Specifically, the following Theorem, which adapts Theorem 3 in Ward and Glynn [84] to our setting, shows that Theorem 2.1 continues to hold in the more general \( GI/GI/N + GI \) setting. Although the Theorem statement assumes \( N = 1 \), we expect that its statement remains valid for any fixed \( N \) that is independent of \( \lambda \) in the conventional HT limit regime. This is because in that regime the number of customers waiting is very
large compared to the number of customers in service, so that the limiting behavior of the number-in-system and number-in-queue processes is identical.

**Theorem 3.1** Let \(Q^\lambda\) denote the number-in-system process for the \(GI/GI/N + GI\) model having parameters \(\lambda, \mu^\lambda\), and \(N = 1\) fixed, given as in the conventional HT limit regime in Definition 2.1, and also \(c^2_a, c^2_s\), and \(F\) fixed. Then,

\[
\frac{Q^\lambda}{\sqrt{\lambda}} \Rightarrow X_{ROU}, \quad \text{as } \lambda \to \infty, \text{ in } D[0, \infty),
\]

where \(X_{ROU}\) is the reflected Ornstein-Uhlenbeck process given in (2.4), with infinitesimal drift function

\[
m_{ROU}(x) = -\beta - F'(0)x,
\]

and infinitesimal variance \(\sigma^2 = c^2_a + c^2_s\).

In words, Theorem 3.1 states that the number-in-system process in the \(GI/GI/1 + GI\) queue has an almost identical limiting behavior to the number-in-system process in the \(M/M/1 + M\) model, except that the value of the reneging density at 0, \(F'(0)\), replaces the reneging rate \(\gamma\) in the limiting diffusion, and the infinitesimal variance now depends on the coefficient of variation of the inter-arrival and service times. Note that Theorem 3.1 is consistent with Theorem 2.1 because when \(F\) is the distribution function associated with an exponential random variable having mean \(1/\theta\), \(F'(0) = \theta\), and the associated squared coefficient of variation is 1, so that for the \(M/M/N + M\) model in Section 2, \(F'(0) = \gamma\), and \(c^2_a = c^2_s = 1\), implying that \(\sigma^2 = 2\).

The role of \(F'(0)\) in Theorem 3.1 is that it governs the state-dependence in the infinitesimal drift function. When we relate back to the \(M/M/N + M\) setting, we see that the parameter \(\gamma\) determines both the state-dependence in the infinitesimal drift function and also the instantaneous rate of customer reneging from the system. In other words, \(\gamma X_{ROU}(t)\), when multiplied by \(\sqrt{\lambda}\), can be thought of as an approximation for the instantaneous rate of customer reneging at time \(t > 0\) in an \(M/M/N + M\) system, \(\gamma [Q - N]^+\), when \(N\) is small. Similarly, the appropriate way to interpret the value of \(F'(0)X_{ROU}(t)\), when multiplied by \(\sqrt{\lambda}\), is as an approximation for the instantaneous customer reneging rate. Then, \(\sqrt{\lambda} \int_0^t F'(0)X_{ROU}(s)ds\) provides an approximation for the cumulative number of customer that have reneged up to time \(t > 0\).

Intuitively, the reason that the value of the reneging time density function at 0 is important is that in the conventional HT limit regime, waiting times are small. This is because the queue length is of order \(\sqrt{\lambda}\), so that Little’s law implies that the waiting times are of order \(\sqrt{\lambda}/\lambda = 1/\sqrt{\lambda}\). Since it is also true that waiting times in the HW limit regime are small, we expect that the value of the abandonment time density function at 0 will appear in the limit diffusion that arises in the HW limit regime as well. This intuition is correct, and we formalize it by showing that the role of \(F'(0)\) in the HW limit regime is exactly the same as in the conventional HT limit regime; that is, it can be used to form an approximation for the instantaneous customer reneging rate.
We isolate the role of $F'(0)$ for the $GI/GI/N + GI$ queue in the HW limit regime in the manner discussed in the preceding paragraph because in general the limit process that arises as an approximation to the centered and scaled number-in-system process for the $GI/GI/N + GI$ queue in this regime is very complicated. This is because the residual service times matter in the limit, so that, consistent with what is known about the $GI/GI/N$ model (see, for example, the discussion at the beginning of Section 5 in [43], or the limit process in Reed [70] [71]), we cannot expect that the limit process is in general Markov. The limit process for the $GI/GI/N + GI$ queue in the HW regime is given in Theorem 1 and Corollary 1 in Mandelbaum and Momcilovic [61], and is, as expected, complicated. The limit process when the service distribution is phase type is somewhat simpler, and is established in Theorem 1 in Dai, He, and Tezcan [28]. The idea here is to state a result that confirms that $F'(0)$ also dictates the instantaneous customer reneging rate in the HW limit regime without having to specify the full limit diffusion (which requires providing a fuller background on the additional complexities that arise in the HW limit regime when service times have general distributions, and does not enhance the understanding of the connection between the effect of customer reneging in the conventional HT and HW limit regimes). This is done in the following theorem; see Theorem 2.1 in Dai and He [27] and Corollary 3 in [61]4.

**Theorem 3.2** Let $Q^\lambda$ denote the number-in-system process and $L^\lambda = \{L^\lambda(t) : t \geq 0\}$ track the cumulative number of customer abandonments in $[0, t]$ in the $GI/GI/N + GI$ model having parameters $\lambda$, $\mu$ fixed, and $N^\lambda$, given as in the HW limit regime in Definition 2.2, and also $c_2^a$ and $F$ fixed. Then, for any $T > 0$,

$$\frac{1}{\sqrt{\lambda}} \sup_{0 \leq t \leq T} \left| L^\lambda(t) - F'(0) \int_0^t \left[ Q^\lambda(s) - N^\lambda \right]^{+} ds \right| \to 0, \text{ in probability, as } \lambda \to \infty.$$ 

Before proceeding to the case that the HW limit regime does give rise to a simple limit diffusion, it should be noted that there are two other proposed approximations for many-server models that have generally distributed service times that do not use diffusions. Iravani and Balcioglu [47] approximate an $M/GI/N + GI$ queue by scaling up the service rate in an $M/GI/1 + GI$ queue. Whitt [89] develops an algorithm to rapidly compute steady-state performance measure approximations in a finite capacity $M/GI/N + GI$ queue that is based on an appropriate finite capacity $M/M/N + M(n)$ queue (that is, an $M/M/N + M$ queue with a state-dependent abandonment rate).

A simple limit diffusion arises in the HW limit regime when service times follow an exponential distribution; that is, for the $GI/M/N + GI$ model. Then, the limit diffusion is the $X_{HW}$ process that solves (2.5) when $\gamma$ is replaced by $F'(0)$ and $\sigma^2 = c_2^a + c_2^s$. This is established in the following Theorem, which is taken from Example 2 in [61], that shows how the simple $X_{HW}$ process arises from the complex limit diffusion in their Theorem 1 for

---

4The result in [27] is a stronger version of Corollary 3 in [61], and holds under more general assumptions than the ones we have made in Section 3.1. However, the statement of the result in [27] also assumes that the sequence $\{Q^\lambda/\sqrt{\lambda}\}$ is stochastically bounded. The stochastic boundedness of the sequence $\{Q^\lambda/\sqrt{\lambda}\}$ is shown in [61].
the $GI/GI/N + GI$ model, in the case that service times are exponential. In particular, we see that Theorem 2.2 continues to hold in the more general $GI/M/N + GI$ setting.

**Theorem 3.3** Let $Q^\lambda$ denote the number-in-system process for the $GI/M/N + GI$ model having parameters $\lambda$, $\mu = 1$ fixed, and $N^\lambda$, given as in the HW regime in Definition 2.2, and also $c^2_a$ and $F$ fixed. Then,

$$\frac{Q^\lambda - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{HW}, \text{ as } \lambda \to \infty, \text{ in } D[0, \infty),$$

where $X_{HW}$ is the piecewise linear diffusion process given in (2.5), with infinitesimal drift function

$$m_{HW}(x) = \begin{cases} -\beta - \mu x & \text{if } x \leq 0 \\ -\beta - F'(0)x & \text{if } x > 0 \end{cases},$$

and infinitesimal variance $\sigma^2 = c^2_a + 1$.

**Remark 3.1** A related model is the $GI/GI/N$ model in which customers balk instead of renege. In the balking model, an arriving customer that has an abandonment time smaller than the time he must wait to receive service never enters the system. Hence all waiting customers in the system eventually receive service, or, equivalently, do not renege. Then, the number-in-system process in the reneging model dominates the number-in-system process in the balking model. Even so, the limiting behavior of the number-in-system process associated with both models is identical, as shown in both [84] and [61]. The point is that, in the conventional HT and HW limit regimes, the two models are indistinguishable.

**Remark 3.2** We have focused on the number-in-system process in the $GI/GI/N + GI$ model. Two other important processes are the offered waiting time process and the observed workload process. The offered waiting time process tracks the amount of time an arriving customer that is infinitely patient would have to wait to receive service. Since customers that eventually renege do not affect the time such a customer would have to wait, the value of the offered waiting time process at time $t$ is exactly the sum of the service times of all customers waiting for service that eventually receive service. The observed workload process tracks the total processing time of all customers in the system at each point in time, including both those that will eventually receive service and those that will eventually abandon. In light of Remark 3.1, it should not be surprising that, under appropriate scaling, these two processes converge to the same limit process in both the conventional HT and the HW limit regimes. Furthermore, there is a Little’s law type relationship that holds between these two processes and the queue-length process. See [84] and [61]. Also see Talreja and Whitt [78] for a two-parameter version of Puhalskii’s invariance principle that shows how to establish

---

5Recall from the third paragraph in the Introduction that we are focusing on models in which the time at which each individual customer will renege is unknown to the system manager. In the case that such times are known, the observed workload process would not include the service times of those customers that eventually abandon.
stochastic-process limits for the offered waiting time process given stochastic process limits for the arrival, queue-length, departure, and abandonment processes.

Remark 3.3 As a follow-up to Remark 2.2, in the case that customers may renege while in service, Theorem 3.1 holds as stated, and Theorem 3.3 is no longer valid. We are not aware of any results in the literature that cover GI/GI/N + GI models in the HW limit regime, in which customers may renege while in service.

Theorems 3.1 and 3.3 rigorously validate the approximations (2.6) and (2.7) in Section 2.3 based on \( X_{ROU} \) and \( X_{HW} \), when \( \gamma \) is replaced by \( F'(0) \) and \( \sigma^2 = c_a^2 + c_s^2 \). This suggests that the steady-state approximations in (2.9) and (2.10) are also valid. The steady-state approximation in (2.9) is supported by numeric computation in [84], and certain steady-state performance measure approximations are rigorously justified by Theorem 4.1 in Mandelbaum and Zeltyn [63] for an \( M/M/N + GI \) model, although not the full steady-state distribution approximation for \( Q \). However, neither [84] nor [61] provide the rigorous argument.

Performance measure approximations based on \( X_{ROU} \) and \( X_{HW} \) can be refined by considering corrected diffusion approximations. Recent work for models without abandonment includes Blanchet and Glynn [18] under conventional HT scaling and Janssen, van Leeuwaarden and Zwart [48] under HW scaling. An interesting direction for future research is to develop corrected diffusion approximations for models with abandonment.

3.3 Hazard Rate Scaling

One natural question that arises when using Theorems 3.1 and 3.3 to develop performance measure approximations is: what do we do when \( F'(0) = 0 \), or when \( F'(0) \to \infty \) as \( x \downarrow 0 \)? For example, the gamma density function has value 0 at the origin when its shape parameter exceeds 1 and is \( \infty \) at the origin when its shape parameter is smaller than 1 (and is an exponential distribution when its shape parameter equals 1). Hence we are motivated to refine the results in Theorems 3.1 and 3.3 to incorporate such situations. To do this, we focus our attention on the hazard rate function associated with the abandonment distribution.

We begin by intuitively motivating how we expect the hazard rate function to appear in the limit diffusion. First, as in (3.3) in [89], an approximate Markovian abandonment rate for the \( j \)th customer from the end of the queue is \( h(j/\lambda) \), where \( h \) is the hazard rate function associated with \( F \). This is because when abandonments are relatively rare, we expect that close to \( j-1 \) customers have arrived before the \( j \)th, so that the \( j \)th customer has been waiting approximately \( j/\lambda \) time units. Then, the total instantaneous abandonment rate from the queue is \( \sum_{j=1}^{Q^\lambda(t)} h(j/\lambda) \). Theorems 3.1 and 3.3 suggest scaling the instantaneous customer abandonment rate by \( \sqrt{\lambda} \). Next, if we also allow \( h \) to depend on \( \lambda \), so that \( h^\lambda(x) = h(\sqrt{\lambda}x) \) for all \( x \geq 0 \), the scaled instantaneous abandonment rate becomes

\[
\frac{1}{\sqrt{\lambda}} \sum_{j=1}^{Q^\lambda(t)} h^\lambda(j/\lambda) = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{\sqrt{\lambda}(Q^\lambda(t)/\sqrt{\lambda})} \sqrt{\lambda} h(j/\sqrt{\lambda}). \tag{3.1}
\]
When $Q^{\lambda}/\sqrt{\lambda}$ weakly converges to a limit process $\hat{Q}$, (3.1) suggests that
\[
\frac{1}{\sqrt{\lambda}} \sum_{j=1}^{Q^{\lambda}(t)} h^{\lambda}(j/\lambda) \Rightarrow \int_0^{\hat{Q}(t)} h(x)dx, \text{ as } \lambda \to \infty.
\] (3.2)

Then, (3.2) suggests that $X_{ROU}$ in Theorem 3.1 and $X_{HW}$ in Theorem 3.3 should be replaced by the diffusions $X_{ROU-HS}$ and $X_{HW-HS}$ (where HS is mnemonic for hazard scaling) having infinitesimal drift functions
\[
m_{ROU-HS}(x) = -\beta - \int_0^x h(y)dy \text{ for } x \geq 0 \text{ and } m_{HW-HS}(x) = \begin{cases} 
-\beta - \mu x & \text{if } x \leq 0 \\
-\beta - \int_0^x h(y)dy & \text{if } x > 0
\end{cases},
\]
and having the same infinitesimal variance $c_a^2 + c_s^2$.

The last sentence in the previous paragraph is exactly true. This is seen in the following Theorem, which combines Theorem 5.1 part (i) and Theorem 6.1 in Reed and Ward [73] and Theorem 3.1 in Reed and Tezcan [72], and adapts them to our setting.

**Theorem 3.4** Assume that $F$ is absolutely continuous with hazard rate function $h$ that is bounded.

(i) Under the conditions of Theorem 3.1,
\[
\frac{Q^{\lambda}}{\sqrt{\lambda}} \Rightarrow X_{ROU-HS}, \text{ as } \lambda \to \infty, \text{ in } D[0, \infty).
\]

(ii) Under the conditions of Theorem 3.3,
\[
\frac{Q^{\lambda} - N^{\lambda}}{\sqrt{\lambda}} \Rightarrow X_{HW-HS}, \text{ as } \lambda \to \infty, \text{ in } D[0, \infty).
\]

We note that Lee and Weerasinghe [57] have proved a weak convergence result in the setting of part (i) of Theorem 3.4 when the arrival process can be state-dependent.

Theorem 3.4 suggests that the approximations in (2.6) and (2.7) can be improved by replacing $X_{ROU}$ and $X_{HW}$ with $X_{ROU-HS}$ and $X_{HW-HS}$. The associated steady-state approximations formed by replacing $X_{ROU}(\infty)$ in (2.9) by $X_{ROU-HS}(\infty)$ and $X_{HW}(\infty)$ in (2.10) by $X_{HW-HS}(\infty)$ is numerically supported by the computations in [73] and [72], but not rigorously justified. The steady-state distributions of $X_{ROU-HS}(\infty)$ and $X_{HW-HS}(\infty)$ have a relatively simple form, and are given in Proposition 6.1 in [73] and Proposition 3.2 in [72]. We are not aware of any transient performance measure computations for $X_{ROU-HW}$ and $X_{HW-HS}$.

The final question concerns when the improvement that results from using $X_{ROU-HS}$ and $X_{HW-HS}$ to develop performance measure approximations will be significant. As in Section
4 in [72], this can best be answered by performing a Taylor series expansion about 0 in the infinitesimal drift of the limit diffusion process as follows

\[-\beta - \int_0^x h(y)dy = -\beta - \left[ h(0)x + \frac{1}{2} h'(0)x^2 + \sum_{j=3}^{\infty} h^{(j)}(0) \frac{x^{j+1}}{(j+1)!} \right]. \tag{3.3}\]

The expansion in (3.3) suggests that \(X_{\text{ROU-HS}} - X_{\text{HS}}\) and \(X_{\text{HW-HS}} - X_{\text{HS}}\) result in significantly better performance measure estimates when \(h'(0)\) is large, since \(h(0) = F'(0)\). The equality (3.3) also justifies the use of the term “refinement” in the first paragraph of this subsection, because (3.3) shows the sense in which Theorem 3.4 refines Theorems 3.1 and 3.3.

**Remark 3.4** In the case that customer reneging times are constant, the FIFO and EDF service disciplines are identical. Under the EDF service discipline, and in conventional HT, the weak convergence of \(Q_\lambda / \sqrt{\lambda}\) follows from the paper [56], which provides the heavy traffic limit for the measure-valued process that tracks the time remaining until each customer in queue will renege (noting that the EDF service discipline can only be used when customer reneging times are known). We re-state their result on the weak convergence of \(Q_\lambda / \sqrt{\lambda}\) in our setting, and then verify that Theorem 3.4 (i) is consistent with their result in the case that reneging times are constant.

Let \(R\) be a two-sided reflected Brownian motion with drift \(-\beta\) and infinitesimal variance \(c_a^2 + c_s^2\) that has state space \([0, b]\) for a given \(b > 0\). Then, if \(Q_\lambda\) is the number-in-system process for the \(GI/GI/N + GI\) model having parameters \(\lambda, \mu, N = 1\) fixed, given as in the conventional HT limit regime in Definition 2.1, and also \(c_a^2, c_s^2,\) and \(F(x) = 1 \{ x \geq b/\sqrt{\lambda} \}\) (so that reneging times are the constant \(b/\sqrt{\lambda}\)), it follows from Theorem 3.5 in [56] that

\[\frac{Q_\lambda}{\sqrt{\lambda}} \Rightarrow R, \text{ as } \lambda \rightarrow \infty, \text{ in } D[0, \infty).\]

Also, when reneging times are the constant \(b/\sqrt{\lambda}\), \(h(x) = 0\) for \(x < b/\sqrt{\lambda}\), so that \(h^\lambda(x) = 0\) for \(x < b\) for all \(\lambda\) and \(m_{\text{ROU-HS}}(x) = -\beta\) for \(x \geq 0\), meaning that the infinitesimal drift of \(X_{\text{ROU-HS}}\) is consistent with that of \(R\). Theorem 5.1 part (ii) in [73] covers the case in which the reneging time distribution has compact support, which then rigorously shows that \(Q_\lambda / \sqrt{\lambda} \Rightarrow R\) as \(\lambda \rightarrow \infty\) under the FIFO service discipline, as well as under the EDF service discipline.

We are not aware of any results for the \(GI/GI/N + GI\) model under the EDF service discipline in the HW regime.

**Remark 3.5** In the case that reneging times are constant, Little’s law suggests that if the reneging times in the system with arrival rate \(\lambda\) are \(b/\sqrt{\lambda}\), then the queue length will rarely exceed \(\lambda \times b/\sqrt{\lambda} = \sqrt{\lambda}b\), so that the \(GI/GI/N + GI\) model behaves similarly to the \(GI/GI/N / \sqrt{\lambda}b\) model (i.e., the \(GI/GI/N\) model with finite capacity \(\sqrt{\lambda}b\)). Theorem 2.1 and Remark 2.2 in Whitt [88] show that if \(Q_\lambda^b\) is the queue-length process associated with the
$GI/GI/N/\sqrt{\lambda b}$ model when $N = 1$, then, for $R$ as given in Remark 3.4,

$$\frac{Q^\lambda_b}{\sqrt{\lambda_b}} \Rightarrow R \text{ as } \lambda \to \infty.$$  

Hence results for reneging models are consistent with results on finite capacity models.

**Remark 3.6** It is convenient in the presentation, but should not be strictly necessary, to assume that inter-arrival times are i.i.d. in order that a version of Theorems 3.1-3.4 hold. What is in general required for Theorem 3.1, 3.3, and 3.4 is that the arrival process obeys a functional central limit theorem. It is necessary to assume that service and reneging times are i.i.d. ([84], [73], and [72] assume i.i.d. inter-arrival times; [61] assume the arrival process obeys a FCLT). Theorem 3.2 holds under even milder assumptions on the arrival process, including when it is time-varying; see [27].

### 3.4 Overloaded Models

We end this section by illuminating one case in which there can be very simple expressions that approximate steady-state performance measures for the $GI/GI/N + GI$ model. In particular, this occurs in overloaded models. An overloaded model arises when the traffic intensity $\lambda/(N\mu)$ remains above 1 as the arrival rate becomes large. This is in contrast to the situation in the conventional HT and HW limit regimes in Definitions 2.1 and 2.2, in which the traffic intensity $\lambda/(\mu^\lambda N) \to 1$, as $\lambda \to \infty$ in conventional heavy traffic, and $\lambda/(\mu N^\lambda) \to 1$ as $\lambda \to \infty$ in the HW regime.

We first heuristically motivate the proposed approximations for the steady-state abandonment probability and the steady-state waiting time given that a customer is served, and then we comment on their rigorous justification. Consider a $GI/GI/N + GI$ model with parameters $\lambda, c_a^2, \mu, c_s^2, N$, and $F$ in which the traffic intensity $\rho = \lambda/(N\mu) > 1$. Under a fluid or law of large numbers scaling, in which we think of customers as fluid flowing through the system continuously, the fluid arrival rate is $\rho$, and the fluid service rate is 1. Then, the rate at which fluid reneges is $\rho - 1$, because there is not enough service capacity. Also, we expect that the system behavior will be dictated by deterministic equations, so that all served fluid spends the same amount of time $w$ in the system. Then, $F(w)$ is the proportion of entering fluid that eventually reneges, so that $\rho F(w)$ is the rate at which customers that eventually reneging enter the system. Equating these two rates shows that

$$\rho F(w) = \rho - 1, \quad (3.4)$$

and, noting that $F(w)$ would be the steady-state abandonment probability if all customers waited $w$ time units implies that

$$P_a \approx \frac{\rho - 1}{\rho}. \quad (3.5)$$

When $F$ is strictly increasing, there is exactly one $w$ that solves (3.4). Then, the value of $w$ that solves (3.4) approximates the expected conditional steady-state waiting time given that a customer is served, which we denote by $W_s$.  

""
The proposed approximations for $W_s$ appear to be identical in both the overloaded regime for a single-server model in Jennings and Reed [49] and in the overloaded many-server regime in Whitt [90]. In the overloaded many-server regime, [90] conjectures the appropriate fluid limit, and shows how the performance measure approximations (3.4) and (3.5) (along with many others) follow from that fluid limit. Theorem 1 in [49] establishes a fluid limit for the offered waiting time process, the process that tracks the amount of time an arriving customer that will eventually be served will have to wait for service, but they do not show the steady-state of that fluid limit in general (in part because the focus of that paper is different). They calculate the steady-state of the fluid limit in exactly two cases:

1. when the abandonment distribution is exponential with mean $1/\gamma$, in which case the steady-state of the fluid limit is $\gamma^{-1} \ln(\rho)$, and

2. when the abandonment distribution is uniform $[a, b]$, in which case the steady-state of the fluid limit is $b - \rho^{-1}(b - a)$.

The expressions in (1) and (2) exactly coincide with the $w$ obtained through solving (3.4) for $w$, under the assumptions that $F$ represents the exponential and uniform distributions respectively.

Another approach is to scale the steady-state performance measures directly, and calculate their limiting values (instead of first showing convergence to the fluid limit, and then finding the steady-state of that fluid limit). This is done in Bassamboo and Randhawa [15] for an $M/M/N + GI$ model. It turns out that the customer abandonment probability approximation is extremely accurate. In particular, in both the conventional HT and HW limit regimes, there exist finite constants $c_1, c_2 > 0$ such that

\[
\frac{\rho - 1}{\rho} \leq P^\lambda_a \leq \frac{\rho - 1}{\rho} + c_1 e^{-c_2 \lambda} \quad \text{for all } \lambda;
\]

see Theorems 2 and 4 in [15], and note that in the notation in their paper $P^\lambda_a$ is $\alpha \lambda / \lambda$. It is also true that the fluid approximation for the number of customers waiting to be served is extremely accurate (up to $O(1)$); see Theorems 1 and 4 in [15].

The story in the overloaded many-server regime is complete. Kang and Ramanan [53], building on the work of Kaspi and Ramanan [55], rigorously establish the fluid limit conjectured in [90], and Kang and Ramanan [52] provide conditions under which the steady-state distributions converge. They do this by representing the state of the system in terms of measure valued processes, including one process that tracks the waiting times of the customers in queue and one process that tracks the amount of time each customer being served has been in service. We expect that one way to complete the story in the overloaded regime for a model with a fixed number of servers is to use a measure valued process that tracks the waiting times of the customers in queue, although it should not be necessary to track the amount of time a customer being served has been in service.
4 A Unifying Diffusion Approximation

Instead of proposing several diffusion approximations for the \( GI/M/N + GI \) model, we would prefer to propose only one. In particular, it is of interest to have one unifying diffusion that in some sense “knits together” the \( X_{ROU} \) and \( X_{HW} \) approximations in (2.6) - (2.8), in the same spirit as the universal diffusion approximation proposed in [82]. The key to this is to find a diffusion process that has the same limiting behavior as \( Q \) in both the conventional HT and HW limit regimes, and also in the intermediate limit regime. This idea is in the same spirit as the universal diffusion approximations discussed in Section 3 (in the context of an \( M/M/1 \) queue) and Section 5 (in the context of an \( M/M/1 + M \) queue) in [82].

We form the proposed unifying diffusion approximation \( X \) by matching its infinitesimal mean to that of \( Q \) when \( Q \) is a continuous-time Markov chain as in Section 2, judiciously choosing its infinitesimal variance, and replacing the discrete state space of \( Q \), \( \{0, 1, 2, \ldots \} \) with the continuous state space \([0, \infty)\). Then, the infinitesimal mean of \( X \) is

\[
 m_X(x) := m_Q(x) = \lambda - N\mu + \mu [N-x]^+ - F'(0) [x-N]^+ ,
\]

assuming that the reneging distribution \( F \) is a proper distribution for which \( F'(0) > 0 \) exists and is finite. We do not match the infinitesimal variances because the infinitesimal variance of \( Q \) is state-dependent, and a state-dependent infinitesimal variance for \( X \) would lead to a process that is less convenient analytically\(^6\). Instead, we let the infinitesimal variance for \( X \) be \( \lambda(c_a^2 + c_s^2) \), for all \( x \geq 0 \), a constant that is consistent with the behavior of the infinitesimal variance of \( Q \) in all three limit regimes. Then, the proposed unifying diffusion \( X \) is a piecewise linear diffusion that solves the stochastic equation

\[
 X(t) = X(0) + \int_0^t m_X(X(s)) ds + \sqrt{\lambda(c_a^2 + c_s^2)} B(t) + L(t), \tag{4.1}
\]

where \( X(0) \geq 0 \), \( B \) is a standard Brownian motion, and \( L \) is the minimal non-decreasing process under which \( X(t) \geq 0 \) for all \( t \geq 0 \).

In summary, given the \( GI/M/N + GI \) model with associated parameters \( \lambda, c_a, \mu, c_s, N, \) and \( F \), we propose the (unscaled) approximation

\[
 Q(\cdot) \overset{D}{\approx} X(\cdot), \tag{4.2}
\]

under the assumption that \( Q(0) = X(0) \). The analytic analysis that justifies the approximation (4.2) is that the asymptotic behavior of \( Q \) and \( X \), under appropriate scaling is the same in both the conventional HT and HW limit regimes. In particular, the following theorem, when combined with either Theorems 2.1 and 2.2 (in the case of an \( M/M/N + M \) queue) or Theorems 3.1 and 3.3 (in the case of a \( GI/M/N + GI \) queue), shows that the distributions of \( Q \) and \( X \) are close in both of these regimes (which may at first seem surprising given that \( Q \) is a non-negative integer-valued process and \( X \) is a non-negative continuous process).

\(^6\)In particular, although the problem of finding an expression for the process’s steady-state distribution should be tractable, because it is a one-dimensional diffusion, transient analysis of the process may be much more unwieldy.
Theorem 4.1 Let $\beta \in \mathbb{R}$.

1. (Conventional HT Regime) Let $X^\lambda$ be the piecewise linear diffusion that solves (4.1) and has parameters $\lambda, \mu^\lambda$, and $N$, given as in the conventional HT limit regime in Definition 2.1, and $c_a^2, c_s^2, F$ fixed. Then,
   \[
   \frac{X^\lambda}{\sqrt{\lambda}} \Rightarrow X_{ROU}, \text{ as } \lambda \to \infty, \text{ in } D[0, \infty).
   \]

2. (HW Many-Server Regime) Let $X^\lambda$ be the piecewise linear diffusion that solves (4.1) and has parameters $\lambda, \mu = 1$, and $N^\lambda$ given as in the HW limit regime in Definition 2.2, and $c_a^2, c_s^2$, and $F$ fixed. Then,
   \[
   \frac{X^\lambda - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{HW}, \text{ as } \lambda \to \infty, \text{ in } D[0, \infty).
   \]

The implication of Theorem 4.1 is that if $d$ is the Prohorov metric (the metric that gives rise to the topology of weak convergence on $D[0, \infty)$; see, for example, Ethier and Kurtz [29]), then
\[
d\left(P\left(\frac{Q^\lambda}{\sqrt{\lambda}} \in \cdot \right), P\left(\frac{X^\lambda}{\sqrt{\lambda}} \in \cdot \right)\right) \to 0, \text{ as } \lambda \to \infty,
\]
in the conventional HT limit regime, and
\[
d\left(P\left(\frac{Q^\lambda - N^\lambda}{\sqrt{\lambda}} \in \cdot \right), P\left(\frac{X^\lambda - N^\lambda}{\sqrt{\lambda}} \in \cdot \right)\right) \to 0, \text{ as } \lambda \to \infty,
\]
in the HW limit regime.

**Proof of Theorem 4.1:** Define
\[
\hat{X}^\lambda_1 := \frac{X^\lambda}{\sqrt{\lambda}} \text{ and } \hat{X}^\lambda_2 := \frac{X^\lambda - N^\lambda}{\sqrt{\lambda}}.
\]

The weak convergence $\hat{X}^\lambda_1 \Rightarrow X_{ROU}$ as $\lambda \to \infty$ in part 1 can be proved by first noting that the infinitesimal means and variances converge in the conventional HT regime, since
\[
\lim_{h \downarrow 0} \frac{1}{h} E \left[ \hat{X}^\lambda_1(h) - \hat{X}^\lambda_1(0) | \hat{X}^\lambda(0) = x \right] = \frac{1}{\sqrt{\lambda}} m_X(\sqrt{\lambda}x) \to -\beta - F'(0)x = m_{ROU}(x), \text{ as } \lambda \to \infty,
\]
and
\[
\lim_{h \downarrow 0} \frac{1}{h} E \left[ \left( \hat{X}^\lambda_1(h) - \hat{X}^\lambda_1(0) \right)^2 | \hat{X}^\lambda(0) = x \right] = \frac{1}{\lambda} \times \lambda(c_a^2 + c_s^2) = (c_a^2 + c_s^2).
\]

The next step is to parallel the arguments used to prove Theorem 2 part (i) in [82], which appeals to semigroup theory to handle the reflection term.
The weak convergence $\hat{X}_2^\lambda \Rightarrow X_{HW}$ as $\lambda \to \infty$ in part 2 follows from Stone’s theorem [76], because the infinitesimal means and variances converge in the HW limit regime as follows

$$\lim_{h \downarrow 0} \frac{1}{h} E \left[ \left( \hat{X}_2^\lambda(h) - \hat{X}_2^\lambda(0) \right)^2 | \hat{X}_2^\lambda(0) = x \right] = \frac{1}{\lambda} \times \lambda (c_a^2 + c_s^2) = (c_a^2 + c_s^2).$$

We also conjecture that if $X^\lambda$ is the piecewise linear diffusion that solves (4.1) and has parameters $\lambda$, $\mu^\lambda$, and $N^\lambda$, given as in the intermediate limit regime in Definition 2.3, and $c_a^2$, $c_s^2$, and $F$ fixed, then

$$\frac{X^\lambda - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{ROU}, \text{ as } \lambda \to \infty, \text{ in } D[0, \infty), \quad (4.3)$$

so that

$$d \left( P \left( \frac{Q^\lambda - N^\lambda}{\sqrt{\lambda}} \in \cdot \right), P \left( \frac{X^\lambda - N^\lambda}{\sqrt{\lambda}} \in \cdot \right) \right) \to 0, \text{ as } \lambda \to \infty,$$

in the intermediate limit regime as well as in the conventional HT and HW limit regimes. Intuitively (4.3) follows because

$$\frac{1}{\sqrt{\lambda}} m_X \left( \sqrt{\lambda x} + N^\lambda \right) = -\beta + \mu \left( \lambda^{1-a} + \lambda^{1/2-a} \beta \right) [-x]^+ - F'(0)[x]^+ \to \begin{cases} -\beta - \gamma x & \text{if } x \geq 0 \\ -\infty & \text{if } x < 0 \end{cases},$$

which is consistent with the limiting behavior of $\frac{1}{\sqrt{\lambda}} m_Q \left( \sqrt{\lambda x} + N^\lambda \right)$, and the calculation

$$\lim_{h \downarrow 0} \frac{1}{h} E \left[ \left( \hat{X}_2^\lambda(h) - \hat{X}_2^\lambda(0) \right)^2 | \hat{X}_2^\lambda(0) = x \right] = (c_a^2 + c_s^2)$$

remains valid.

Finally, we perform a small numeric study that investigates the accuracy of the approximation (4.2). We assume a $M/M/N + M$ model, because in that situation exact performance measure computations are available, as shown in Section 2.1. Since our objective was to have a diffusion approximation that in some sense unifies the conventional HT and HW limit regimes, it is natural to fix $\lambda$ and the traffic intensity $\rho := \lambda/(N\mu)$, and to vary the number of agents $N$ and the service speed $\mu$, so that the relationship $\lambda = N\mu\rho$ is preserved. In Table 4.1, we let $\lambda = 50$, $\gamma = 1$, fix $\rho = 1$ (equivalently $N\mu = 50$), and calculate $P_a$, \footnote{This choice of parameters coincides with Graph b in Appendix A in [33] when $N = 50$.}
Table 4.1: Exact and approximated performance measure calculations for an $M/M/N+M$ model having $\lambda = 50$, $\gamma = 1$, $\lambda/(N\mu) = 1$, and varying values of $N$ and $\mu$.

$E[Q(\infty)]$, and $E[X(\infty)]$. We use the analysis tool 4CallCenters [59] to find $P_a$ and $E[Q(\infty)]$, and compute $E[X(\infty)]$ according to the formulae in [24], which yields

$$E[X(\infty)|X(\infty) \leq N] = \frac{\lambda}{\mu} + \sqrt{\frac{\lambda}{\mu}} \frac{\phi\left(-\sqrt{\frac{\lambda}{X \mu}}\right) - \phi\left(\sqrt{\frac{\lambda}{X}} \left(N - \frac{1}{\mu}\right)\right)}{\Phi\left(\sqrt{\frac{\lambda}{X}} \left(N - \frac{1}{\mu}\right)\right) - \Phi\left(-\sqrt{\frac{\lambda}{X \mu}}\right)}$$

$$E[X(\infty)|X(\infty) > N] = m + \sqrt{\frac{\lambda}{\gamma}} \frac{\phi\left(\sqrt{\frac{\lambda}{X}} (N - m)\right)}{1 - \Phi\left(\sqrt{\frac{\lambda}{X}} (N - m)\right)},$$

where $m := (\lambda - N\mu + \gamma N)/\gamma$, and

$$P(X(\infty) \leq N) = \left[1 + \sqrt{\frac{\mu}{\gamma}} \frac{\phi\left(\sqrt{\frac{\lambda}{X}} (N - \frac{1}{\mu})\right)}{\Phi\left(\sqrt{\frac{\lambda}{X}} (N - \frac{1}{\mu})\right) - \Phi\left(-\sqrt{\frac{\lambda}{X \mu}}\right)}\right]^{-1},$$

recalling that $\phi$ and $\Phi$ represent the standard normal pdf and cdf respectively, so that

$$E[X(\infty)] = E[X(\infty)|X(\infty) \leq N] P(X(\infty) \leq N) + E[X(\infty)|X(\infty) > N] P(X(\infty) > N).$$

For comparison purposes, in Table 4.1, we also display the results of using the approximations supported by Theorems 2.1 and 2.2. In both the conventional HT and HW limit regimes, $N\mu = \lambda + \beta \sqrt{\lambda}$, which implies $\beta = 0$ for all values of $N$ and $\mu$ in Table 4.1. Then, from Theorem 3.1 and approximation (2.9)

$$E[Q(\infty)] = \sqrt{\lambda} E[X_{ROU}(\infty)] = \sqrt{\lambda} \frac{\phi(0)}{1 - \Phi(0)} = 5.64,$$

and from Theorem 3.3 and approximation (2.10)

$$E[Q(\infty)] = \sqrt{\lambda} E[X_{HW}(\infty)] + N$$

$$= \sqrt{\lambda} \left(\frac{1}{\sqrt{\mu} \Phi(0)} \left(\frac{1}{1 + \sqrt{\mu} / \sqrt{\gamma}}\right) + \frac{1}{\sqrt{\gamma} \Phi(0)} \left(\frac{\sqrt{\mu} / \sqrt{\gamma}}{1 + \sqrt{\mu} / \sqrt{\gamma}}\right)\right) + N.$$
Table 4.1 shows that the approximation (4.2) estimates the steady-state mean number-in-system extremely accurately. As expected, the conventional HT approximation in (2.9) is very accurate (having 2.8% relative error) when $N = 1$, but not very accurate as $N$ grows larger. On the other hand, the HW approximation in (2.10) is very accurate for all values of $N$, even though in general there is no reason to expect it to perform well for small values of $N$ (like $N = 1$). This can raise the question of whether or not there is a need for a single unifying approximation. The answer is “yes”, because it is only the unifying approximation proposed in (4.2) for which there is theoretic justification that it performs well both when $\lambda$ and $\mu$ are large but $N$ is small, and when $\lambda$ and $N$ are large but $\mu$ is small. Furthermore, it is only the unifying approximation that has the potential to also consolidate the H-W and intermediate limit regimes. However, we would ideally like to have a unifying diffusion approximation that is also consistent with the hazard rate scaling in Section 3.3, and the overloaded regimes in Section 3.4. We view the question of finding a better unifying diffusion approximation as an interesting open question for future research.

Acknowledgements

We would like to thank Baris Ata for suggesting this project. We would like to thank Ananda Weerasinghe for helpful comments on an early version.

References


34