On the steady-state probability of delay and large negative deviations for the $GI/GI/n$ queue in the Halfin-Whitt regime

David A. Goldberg
Georgia Institute of Technology, dgoldberg9@isye.gatech.edu, http://www2.isye.gatech.edu/ dgoldberg9/

We consider the FCFS $GI/GI/n$ queue in the Halfin-Whitt heavy traffic regime, and prove bounds for the steady-state probability of delay (s.s.p.d.) for generally distributed processing times. We prove that there exists $\epsilon > 0$, depending on the inter-arrival and processing time distributions, such that the s.s.p.d. is bounded from above by $\exp(-\epsilon B^2)$ as the associated excess parameter $B \to \infty$; and by $1 - \epsilon B$ as $B \to 0$. We also prove that the tail of the steady-state number of idle servers has a Gaussian decay, and use known results to show that our bounds are tight (in an appropriate sense).

Our main proof technique is the derivation of new stochastic comparison bounds for the FCFS $GI/GI/n$ queue, which are of a structural nature, hold for all $n$ and times $t$, and build on the recent work of Gamarnik and Goldberg [21].

Key words: many-server queues, Halfin-Whitt regime, probability of delay, stochastic comparison, weak convergence, large deviations, Gaussian process, renewal process

MSC2000 subject classification: 60K25

1. Introduction
Parallel server queueing systems can operate in a variety of regimes that balance between efficiency and quality of offered service. This is captured by the so-called Halfin-Whitt (H-W) heavy traffic regime. Although studied originally by Erlang [19] and Jagerman [32], the regime was formally introduced by Halfin and Whitt [28], who studied the $GI/M/n$ system (for large $n$) when the traffic intensity $\rho$ scales like $1 - Bn^{-\frac{1}{2}}$ for some strictly positive excess parameter $B$. The authors prove that for this sequence of $GI/M/n$ queueing models (indexed by $n$), the steady-state probability that an arriving job has to wait for service (i.e. steady-state probability of delay) converges (as $n \to \infty$) to a function of $B$ (independent of $n$), which they explicitly compute. This limiting probability converges to 0 as $B \to \infty$ (low-utilization regime), converges to 1 as $B \to 0$ (high-utilization regime), and decreases monotonically from 1 to 0 as $B$ increases from 0 to $\infty$, thus nicely quantifying the trade-off between server utilization and quality of service (as measured by s.s.p.d.). Analogous explicit formulas have been found for the case in which processing times are a mixture of an exponential distribution and a point mass at 0 (i.e. $H^*_2$ distributed) by Whitt [61]; and for the case of deterministic processing times by Jelenkovic et al. [34]. For the case of exponentially distributed processing times, these results have also been extended to allow for abandonments, see Garnett et al. [24] and Mandelbaum and Zeltyn [42].

However, much less is known for more general processing time distributions. This is particularly unfortunate, as it is believed that in many of the major application domains of the H-W regime (e.g. call centers), processing times have more general distributions (e.g. log-normal), see Brown et al. [7]. In this more general setting, the main known results with regards to (w.r.t.) the s.s.p.d. can be summarized as follows. For processing times with finite support, Gamarnik and Momcilovic [22] give an implicit description (in terms of a certain Markov chain) of the limiting s.s.p.d in the
H-W regime (also proving that this limit exists), and show that this probability lies strictly in (0, 1). For hyper-exponentially distributed processing times, Zheng et al. [63] give several bounds for the s.s.p.d. in the H-W regime, under various additional assumptions. Whitt [60] gives several heuristic approximations for this limiting s.s.p.d. for generally distributed processing times, which he verifies numerically. Dai et al. [15] prove the validity of a certain diffusion approximation for the steady-state number in system for the case of phase-type processing times with abandonments, and the work of Dai and He [12] provides numerical algorithms for evaluating the associated probabilities.

A similar situation exists for questions related to the large deviations (l.d.) properties of the steady-state number of idle servers in the H-W regime. Indeed, the settings in which this limiting l.d. behavior is precisely understood coincide exactly with the settings in which the limiting s.s.p.d. can be computed explicitly, as described above. Furthermore, very little is known about even the qualitative behavior / crude asymptotic scaling of either quantity. However, even from the special cases of $GI/H^*_2/n$ and $GI/D/n$ queues for which the relevant limits can be computed in closed form, some interesting patterns emerge. We now briefly describe these patterns heuristically before introducing all relevant model details, as a more complete discussion follows in Section 8. Let $p_B$ denote the limiting s.s.p.d. when the excess parameter is $B$, and $F_{idle,B}(x)$ denote the limiting steady-state probability of seeing more than $xn^{2\epsilon}$ idle servers in the $n$th system under the H-W scaling, where the particular inter-arrival and processing time distributions are to be inferred from context. Note that as $B \to 0$, one would expect $p_B$ to converge to 1, since as $B \to 0$ the system becomes more overloaded. Similarly, as $B \to \infty$, one would expect $p_B$ to converge to 0. We have already asserted that Halfin and Whitt [28] proved this for the case of exponentially distributed processing times. However, one can ask how quickly these quantities approach 1 as $B \to 0$, and 0 as $B \to \infty$. Such qualitative information would provide valuable insight into the s.s.p.d. in the H-W regime, beyond the initial insight of Halfin and Whitt that the correct limits should be 1 and 0. We believe such insight to be especially relevant for queues in the H-W regime, since one of the primary motivating features of the H-W regime was that the s.s.p.d. should have a non-trivial limit as $n \to \infty$.

Then the work of Halfin and Whitt [28], Whitt [61], and Jelenkovic et al. [34] imply the following.

Observation 1. When the processing times are exponentially distributed, $H^*_2$, or deterministic, and the inter-arrival distribution has finite third moment, there exist strictly positive finite constants $\epsilon_1, \epsilon_2, \epsilon_3$, depending only on the inter-arrival and processing time distributions (but not $B$ or $n$), such that (s.t.)

(i) $\lim_{B \to \infty} B^{-2}\log(p_B) = -\epsilon_1$,
(ii) $\lim_{B \to 0} B^{-1}(1 - p_B) = \epsilon_2$,
(iii) $\lim_{x \to \infty} x^{-2}\log(F_{idle,B}(x)) = -\epsilon_3$.

As previously mentioned, it is believed that for the relevant systems which arise in practice, the service time distributions are in fact closer to a log-normal distribution, see Brown et al. [7]. In this paper, we thus set out to answer the following question.

Question 1. To what extent do the scalings suggested by Observation 1 hold in general?

We note that the scaling of (i) and (ii) represent “limits within limits”, where letting $B \to 0$ is moving the H-W scaling “closer” to the classical heavy-traffic scaling in which $\rho$ scales like $1 - \frac{c}{n}$ for some $c > 0$ as $n \to \infty$, while letting $B \to \infty$ is moving the H-W scaling closer to the classical heavy-traffic scaling in which $\rho$ is bounded away from 1 as $n \to \infty$, and we refer the reader to Halfin and Whitt [28] for a more detailed discussion. We note that the regime in which $B \to 0$ has been explicitly discussed in the context of $GI/D/n$ queues by Jelenkovic et al. [34], and Janssen
and Van Leeuwaarden [33]. We further note that if one wishes to bound the s.s.p.d. from below, or the number of idle servers from above (as done by Gamarnik and Stolyar [23] Theorem 4.(i), and Lemmas 5 and 10), one can derive non-trivial bounds in the H-W regime by comparing the true system to an appropriate infinite-server queue, even for the case of generally distributed processing times, and we include a more complete discussion in Section 8. However, it seems that no similarly straightforward comparison yields meaningful bounds in the other direction, i.e. upper bounds on the s.s.p.d. and lower bounds on the number of idle servers, in the H.W. regime. In this paper, we will focus on exactly these settings where straightforward infinite-server bounds seem to break down.

We now briefly review the more general question of when the limits \( p_B \) and \( F_{\text{idle,B}}(x) \) are known to exist, i.e. when the appropriately scaled sequence of queues under the H-W scaling has a weak limit, irregardless of the more subtle question of whether those limits themselves satisfy (i) - (iii). As previously discussed, the relevant weak limits are proven to exist, and actually computed explicitly, for the case of exponentially distributed processing times by Halfin and Whitt [28], \( H_2^* \) processing times by Whitt [61], and deterministic processing times by Jelenkovic et al. [34]. Similar, albeit often less explicit, weak convergence results under the H-W scaling were subsequently obtained for more general multi-server systems by Puhalski and Reiman [48], Mandelbaum and Momcilovic [40], Gamarnik and Momcilovic [22], Gamarnik and Stolyar [23], Kaspi and Ramanan [36], Reed [50], and Puhalski and Reed [49]. Many of these results have also been extended to the setting of abandonments by Mandelbaum and Momcilovic [41], Dai et al. [14], Reed and Tezcan [51], Huang et al. [30], Dai et al. [15], Kang and Ramanan [35], Weerasinghe [57] and we refer the interested reader to the recent surveys of Dai and He [13] and Ward [56].

Unfortunately, as the theory of weak convergence generally relies heavily on the assumption of compact time intervals, the most general of these results hold only in the transient regime. Indeed, the only settings in which the relevant sequence of normalized steady-state queue-length distributions have also been shown to have a weak limit are subsumed by the settings in which the processing time distribution either has finite support, or is of phase-type. We note that although any distribution can be approximated to within any accuracy by either of these families, it is not known how the quality of such an approximation (if its complexity cannot grow with \( n \)) holds up under the H-W scaling. Furthermore, such an approximation may not capture certain features of interest, e.g. large deviations behavior. Furthermore, we observe the following.

**Observation 2** In spite of the extensive body of work on queues in the H-W regime, the answer to Question 1 is unknown beyond the cases of exponentially distributed, \( H_2^* \), and deterministic processing times.

Recently, Gamarnik and Goldberg [21] developed a novel stochastic comparison framework to address some of the issues faced by previous techniques, namely the difficulty of proving results in the steady-state, and the non-explicit descriptions of relevant limits. In that work, the authors analyzed the steady-state \( GI/GI/n \) queue by comparing to a “modified system” in which all servers are kept busy at all times by adding “artificial arrivals” to the system whenever a server would otherwise have gone idle. They showed that the steady-state distribution of this modified system has a very simple representation as the supremum of a certain one-dimensional random walk. Using these techniques, they proved tightness of the relevant sequence of normalized steady-state measures, and computed the limiting large deviation behavior for the number of jobs waiting in queue, under quite general assumptions. Unfortunately, the techniques developed by Gamarnik and Goldberg [21] cannot be used to provide upper bounds on the s.s.p.d., or lower bounds on the number of idle servers, as the aforementioned modified system always has all servers busy.

In this paper, we make considerable progress towards resolving Question 1. We prove that under quite general assumptions on the inter-arrival and processing time distributions, e.g. finite third
moment, there exists some strictly positive \( \epsilon_1, \epsilon_2 \) (depending on the inter-arrival and processing time distributions) s.t. the limiting s.s.p.d. is bounded from above by \( \exp \left( -\epsilon_1 B^2 \right) \) as \( B \to \infty \); and the limiting s.s.p.d. is bounded from above by \( 1 - \epsilon_2 B \) as \( B \to 0 \). We also prove that for any fixed inter-arrival distribution, processing time distribution, and parameter \( B \), there exists \( \epsilon > 0 \) s.t. the probability of there being more than \( x n^{\frac{2}{\epsilon}} \) idle servers (in steady-state, for large \( n \)) is bounded from below by \( \exp \left( -\epsilon x^2 \right) \) as \( x \to \infty \). Combined with known results for the the \( M/D/n, M/H_2/n, \) and \( M/GI/\infty \) queues, our results show that the asymptotic scaling suggested by (i) - (iii) is correct, yielding a partial positive answer to Question 1. We note that although our proofs do not demonstrate the existence of the relevant limits, they do show that all relevant lim inf, lim sup, and associated subsequential limits scale in this way, and we refer the reader to the statement of our main results in Section 2, and our comparison to known results in Section 8, for details.

Our main proof technique is the derivation of new and simple bounds for the FCFS \( GI/GI/n \) queue, which significantly extend the stochastic comparison framework developed by Gamarnik and Goldberg [21]. We consider a different modified system, in which we not only keep certain servers busy by adding “artificial arrivals”, but also allow servers to “break down” at a time of our choosing, leaving a number of “working servers” which is strictly less than \( n \). By then keeping only those “working servers” busy at all times, it becomes possible for the number of jobs in this second modified system to go below \( n \), yielding meaningful bounds on the s.s.p.d. and number of idle servers in the H-W regime. Our techniques allow us to analyze many properties of the \( GI/GI/n \) queue in the H-W regime without having to consider the complicated exact dynamics of the \( GI/GI/n \) queue. Our results can also be viewed as a step towards developing a calculus of stochastic-comparison type bounds for parallel server queues, in which one derives bounds by composing structural modifications to a parallel-server queue (e.g. adding jobs, removing servers, adding servers, etc.) over time. Although stochastic comparison techniques have been widely used to study queueing systems, see Stoyan [54], we believe the bounds developed in this paper to be novel, and particularly suited to studying queues in the H-W regime. Another appealing feature of our results is that, as in the work of Gamarnik and Goldberg [21], our bounds are of a structural nature, hold for all \( n \) and all times \( t \geq 0 \), and have intuitive closed-form representations as the suprema of certain natural processes which converge weakly to Gaussian processes.

The rest of the paper proceeds as follows. In Section 2, we present our main results. In Section 3, we establish our general-purpose upper bounds for the queue length in a properly initialized FCFS \( GI/GI/n \) queue. In Section 4, we prove an asymptotic version of our upper bound in the H-W regime. In Section 5, we prove our bounds on the s.s.p.d. as \( B \to \infty \). In Section 6, we prove our bounds on the s.s.p.d. as \( B \to 0 \). In Section 7, we prove our bounds on the large deviations behavior of the steady-state number of idle servers. In Section 8, we compare to previous results from the literature, which show that our bounds are tight in an appropriate sense. In Section 9 we summarize our main results and comment on directions for future research. We include a technical appendix in Section 10.

2. Main Results

We consider the First-Come-First-Serve (FCFS) \( GI/GI/n \) queueing model, in which inter-arrival times are independent and identically distributed (i.i.d.) r.v.s, and processing times are i.i.d. r.v.s.

Let \( A \) and \( S \) denote some fixed r.v.s with non-negative support such that (s.t.) \( \mathbb{E}[A] = \mu_A^{-1} < \infty, \mathbb{E}[S] = \mu_S^{-1} < \infty, \) and \( \mathbb{P}(A = 0) = \mathbb{P}(S = 0) = 0 \). Let \( \sigma_A^2 \) and \( \sigma_S^2 \) denote the variance of \( A \) and \( S \), respectively. Let \( c_A \) and \( c_S \) denote the coefficients of variation (c.v.) of \( A \) and \( S \), respectively, and \( c_A^2 \) and \( c_S^2 \) the corresponding squared coefficients of variation (s.c.v.).

For excess parameter \( B > 0 \) and number of servers \( n \geq 1 \), let \( \lambda_{n,B} \triangleq n - Bn^{\frac{2}{\epsilon}}. \) For \( n \) sufficiently large to ensure \( \lambda_{n,B} > 0 \) (which is assumed throughout), let \( Q_B^n(t) \) denote the number in system
(number in service + number waiting in queue) at time \( t \) in the FCFS \( GI/GI/n \) queue with inter-
arrival times drawn i.i.d. distributed as \( A\lambda_{n,B}^{-1} \) and processing times drawn i.i.d. distributed as \( S \)
(initial conditions will be specified later), independently from the arrival process. Note that this
scaling is analogous to that studied by Halfin and Whitt \[28\], as the traffic intensity in the \( n \)th system is
\( 1 - Bn^{-\frac{1}{2}} \) in both settings. All processes should be assumed right-continuous with left
limits (r.c.l.l.) unless stated otherwise. All empty summations should be evaluated as zero, and all
empty products should be evaluated as one. For clarity of exposition, statements of (in)equality
with probability (w.p.) 1 are not distinguished from statements of (in)equality.

2.1. Main results

Our main results will require two additional sets of assumptions on \( A \) and \( S \). The first set of
assumptions, which we call the H-W assumptions, ensures that \( \{Q^n_B(t), n \geq 1\} \) is in the H-W scaling
regime as \( n \to \infty \). We say that \( A \) and \( S \) satisfy the H-W assumptions iff \( \mu_A = \mu_S \), in which case we
denote this common rate by \( \mu \). The second set of assumptions, which we call the \( T_0 \) assumptions,
is a set of additional technical conditions we require for our main results.

(i) There exists \( \epsilon > 0 \) s.t. \( \mathbb{E}[A^{2+\epsilon}], \mathbb{E}[S^{2+\epsilon}] < \infty \).
(ii) \( c_A^2 + c_S^2 > 0 \). Namely either \( A \) or \( S \) is a non-trivial r.v.
(iii) \( \limsup_{t \to \infty} t^{-1}\mathbb{P}(S \leq t) < \infty \).
(iv) For all sufficiently large \( n \) and all initial conditions, \( Q^n(t) \) converges weakly to a stationary
measure \( Q^n(\infty) \) as \( t \to \infty \), independent of initial conditions.

We refer the interested reader to Gamarnik and Goldberg \[21\] for a discussion of the restrictiveness
and necessity of these assumptions, as that work uses the same set of assumptions.

We now state our main results. We begin by stating our bound on the s.s.p.d. as \( B \to \infty \).

**Theorem 1.** For any fixed \( A \) and \( S \) which satisfy the H-W and \( T_0 \) assumptions,

\[
\limsup_{B \to \infty} -2 \log \left( \limsup_{n \to \infty} \mathbb{P}(Q^n_B(\infty) \geq n) \right) < 0.
\]

In words, Theorem 1 states that there exists \( \epsilon > 0 \), depending only on \( A \) and \( S \), s.t. the limiting
s.s.p.d. is bounded from above by \( \exp(-\epsilon B^2) \) as \( B \to \infty \). We now give our bounds on the s.s.p.d.
as \( B \to 0 \).

**Theorem 2.** For any fixed \( A \) and \( S \) which satisfy the H-W and \( T_0 \) assumptions, s.t. in addition
\( \mathbb{E}[S^3] < \infty \) and \( c_A^2 > 0 \),

\[
\liminf_{B \to 0} \liminf_{n \to \infty} B^{-1} \mathbb{P}(Q^n_B(\infty) < n) > 0.
\]

In words, Theorem 1 states that there exists \( \epsilon > 0 \), depending only on \( A \) and \( S \), s.t. the limiting
steady-state probability that a job does not have to wait for service, i.e. no delay, is bounded from
below by \( \epsilon B \) as \( B \to 0 \). Equivalently, the limiting s.s.p.d. is bounded from above by \( 1 - \epsilon B \) as
\( B \to 0 \). We note that this result is somewhat surprising, since in light of Theorem 1, one might
expect this probability to scale as \( B^2 \) as \( B \to 0 \), since \( 1 - \exp(-\epsilon B^2) \) behaves like \( \epsilon B^2 \) as \( B \to 0 \).

We now state our bounds on the large deviations behavior for the number of idle servers.

**Theorem 3.** For any fixed \( A \) and \( S \) which satisfy the H-W and \( T_0 \) assumptions, s.t. in addition
\( c_A^2 > 0 \), and any fixed \( B > 0 \),

\[
\liminf_{x \to \infty} x^{-2} \log \left( \liminf_{n \to \infty} \mathbb{P} \left( \left( Q^n(\infty) - n \right) n^{-\frac{1}{2}} \leq -x \right) \right) > -\infty.
\]
In words, Theorem 3 states that there exists $\epsilon > 0$, depending only on $A, S$, and $B$, s.t. the tail of the limiting steady-state number of idle servers is bounded from below by $\exp(-\epsilon x^2)$ as $x \to \infty$.

In Section 8, we will show that all of our main results give the correct scaling for the case of $H_\epsilon^i$ or deterministic processing times, as described in Observation 1. We will also show that Theorems 1 and 3 give the correct scaling for a much larger family of queueing systems, i.e. those systems with Poisson arrival processes, using known results for the $M/GI/\infty$ queue. We note that our results cannot be derived using straightforward infinite-server lower bounds, as in all cases our inequalities point in the other direction. Also, as in Gamarnik and Goldberg [21], our results translate into bounds for any weak limits of the associated sequences of r.v.s.

3. Upper Bound

In this section, we prove general upper bounds for the FCFS $GI/GI/n$ queue, when properly initialized. The bounds are valid for all finite $n$, and work in both the transient and steady-state (when it exists) regimes. Although we will later customize these bounds to the H-W regime to prove our main results, we note that the bounds are in no way limited to that regime. Recall that for a non-negative r.v. $X$ with finite mean $\mathbb{E}[X] > 0$, one can define the so-called residual life distribution of $X$, $R(X)$, as follows. Namely, for all $z \geq 0$,

$$P(R(X) > z) = (\mathbb{E}[X])^{-1} \int_z^\infty P(X > y)dy. \quad (1)$$

Recall that associated with a non-negative r.v. $X$, an equilibrium renewal process with renewal distribution $X$ is a counting process in which the first inter-event time is distributed as $R(X)$, and all subsequent inter-event times are drawn i.i.d. distributed as $X$; an ordinary renewal process with renewal distribution $X$ is a counting process in which all inter-event times are drawn i.i.d. distributed as $X$. Let $\{N_i(t), i = 1, \ldots, n\}$ denote a set of $n$ i.i.d. equilibrium renewal processes with renewal distribution $S$. Let $A(t)$ denote an independent equilibrium renewal process with renewal distribution $A$. For $s \in \mathbb{R}^+$, let $V_i^j(s)$ denote the remaining time (at time $s$) until the first renewal to occur after time $s$ in process $N_i(t)$, $i = 1, \ldots, n$. Let $V_i^j(s)$ denote the length of the $(N_i(s) + j)$th renewal interval in process $N_i(t)$, $j \geq 2$, $i = 1, \ldots, n$. Namely, $V_i^j(s)$ is the length of the $(j - 1)$th renewal interval to be initiated in process $N_i(t)$ after time $s$. Similarly, let $U_i^j(s)$ denote the remaining time (at time $s$) until the first renewal to occur after time $s$ in process $A(t)$, and $U_i^j(s)$ denote the length of the $(A(s) + j)$th renewal interval in process $A(t)$, $j \geq 2$.

For $x \in \mathbb{R}^+$, let $A^x(t) \triangleq A(x, x + t)$, $dA^x(t) \triangleq A^x(t) - A^x(t^-)$, and $A^x(s, t) \triangleq A^x(t) - A^x(s)$. Let $N_{i,s}^x(t) \triangleq N_i(x, x + t)$, $dN_{i,s}^x(t) \triangleq N_i^x(t) - N_i^x(t^-)$, $N_{i,s}^x(s, t) \triangleq N_i^x(t) - N_i^x(s)$, $i = 1, \ldots, n$. For $z \in \mathbb{Z}^+$ s.t. $z \leq n$, let $\tau_{x,0}^z \triangleq 0$, and let $\{\tau_{x,k}^z, k \geq 1\}$ denote the sequence of event times in the pooled renewal process $A^x(t) + \sum_{i=1}^z N_{i,s}^x(t)$. For $y \in \mathbb{R}^+$, let

$$\phi(x, y, z) \triangleq \sup_{0 \leq r \leq y} \left( A^x(y - s, y) - \sum_{i=1}^z N_{i,s}^x(y - s, y) \right).$$

Let $V_i^j \triangleq V_i^j(0)$, and $U_i^j \triangleq U_i^j(0)$, $i = 1, \ldots, n$, $j \geq 1$.

For $v \in \mathbb{R}^+$, and $\eta \in \mathbb{Z}^+$ s.t. $\eta \leq n$, let $Q_{\eta}^v$ denote the FCFS $GI/GI/\eta$ queue with inter-arrival times drawn i.i.d. as $A$, processing times drawn i.i.d. as $S$, and the following initial conditions. For $i = 1, \ldots, n$, there is a single job initially being processed on server $i$, with initial processing time $V_i^1(v)$. There are $\phi(0, v, n)$ jobs waiting in queue, and the first inter-arrival time is $U_i^1(v)$. We note that it will become apparent from the analysis of our bounding systems that these are a convenient set of initial conditions. We let $Q_{\eta}^v(t)$ denote the number in system in $Q_{\eta}^v$ at time $t$. We also let $Q \triangleq Q_n^0$, i.e. the $GI/GI/n$ queue with initial conditions s.t. there is a
single job with remaining processing time $V_i^1$ on server $i$, $i = 1, \ldots, n$, the first inter-arrival time is $U^1$, and there are zero jobs waiting in queue. Also, we define $Q(t) \triangleq Q_n^*(t)$.

We now establish an upper bound for $Q(t)$. To build intuition before stating the actual result, we first describe and briefly provide a heuristic motivation and analysis of the relevant bounding systems. In particular, let $\hat{Q}$ denote the following “modified” queueing system. The initial conditions are s.t. there is a single job initially being processed on server $i$, with initial processing time $V_i^1$, $i = 1, \ldots, n$, and there are zero jobs waiting in queue; also, the first inter-arrival time is $U^1$. Service times are drawn i.i.d. distributed as $S$. Furthermore, in this modified system, the arrival process is an “augmentation” of the arrival process of $Q$, where all servers are kept busy at all times by adding an artificial arrival whenever a server would otherwise have gone idle. We note that in the case of Markovian inter-arrival and processing times, this can be equivalently interpreted as putting a reflecting boundary at state $n$. As was proven by Gamarnik and Goldberg [21], the dynamics of $\hat{Q}$ reduce to those of a relatively simple Lindley recursion, i.e. $\hat{Q}$ can be constructed s.t. for all $t \geq 0$, $\hat{Q}(t) = n + \phi(0, t, n)$, the remaining processing time (at time $t$) of the job on server $i$ equals $V_i^1(t), i = 1, \ldots, n$, and the time until the first arrival (after time $t$) equals $U^1(t)$. Furthermore, since $\hat{Q}$ differs from $Q$ only by having extra arrivals, it can be shown that $\hat{Q}(t)$ stochastically dominates $Q(t)$. We note that although this modified system sufficed for the purposes of Gamarnik and Goldberg [21], i.e. to analyze the large deviations properties of the queue length, it will not suffice for our purposes, as in this modified system all servers are always busy, and thus the only obvious bound for the s.s.p.d. is the trivial bound of one. Furthermore, in this modified system there are never any idle servers, and thus similarly it is not clear how to derive meaningful lower bounds for the number of idle servers, other than the trivial bound of 0.

To remedy this, we will consider two further modified systems. In particular, suppose that we fix some $\eta < n$, and let $\hat{Q}_\eta$ be the following modified queueing system. There are $\eta$ servers. The initial conditions are s.t. there is a single job initially being processed on server $i$, with initial processing time $V_i^1, i = 1, \ldots, \eta$, and there are zero jobs waiting in queue; also, the first inter-arrival time is $U^1$. Service times are drawn i.i.d. distributed as $S$, and the arrival process is again an “augmentation” of the arrival process of $Q$, where all $\eta$ servers are kept busy at all times by adding an artificial arrival whenever a server would otherwise have gone idle. In analogy to the dynamics of $\hat{Q}$, $\hat{Q}_\eta$ can be constructed s.t. for all $t \geq 0$, $\hat{Q}_\eta(t) = \eta + \phi(0, t, \eta)$. Furthermore, since $\eta < n$, we can get non-trivial bounds for the s.s.p.d. by computing $\mathbb{P}(\phi(0, t, \eta) \in [0, n - \eta])$. The modified system $\hat{Q}_\eta$ will suffice for parts of our analysis, notably bounding the s.s.p.d. as $B \to \infty$.

However, for the case $B \to 0$, or the case in which we wish to bound the probability of seeing many (e.g. $xt^2$ for $x > 0$) idle servers, this system will not suffice. The problem is that to derive meaningful bounds in these settings, $\eta$ would have to be so small that the system would become unstable or nearly so, making it hopeless to analyze the steady-state. To remedy this, we consider a second modified system. Suppose that we fix some time $\gamma \in [0, t]$, and let $\hat{Q}_{\gamma, \eta}$ be the following second modified queueing system. $\hat{Q}_{\gamma, \eta}$ is identical to $\hat{Q}$ on $[0, \gamma]$ (i.e. $n$ servers with augmented arrival process). However, in $\hat{Q}_{\gamma, \eta}$, at time $\gamma$ servers $\eta + 1, \ldots, n$ “break down”, in the sense that after completing any jobs which they are processing at time $\gamma$, they cannot process any further jobs. Furthermore, on $[\gamma, t]$, servers $1, \ldots, \eta$ are kept busy at all times by adding an artificial arrival whenever one of these servers would otherwise have gone idle, i.e. the arrival process is again a certain augmentation of that seen by $Q$. The dynamics of $\hat{Q}_{\gamma, \eta}$ are again governed by a certain Lindley-type recursion, and $\hat{Q}_{\gamma, \eta}$ can be constructed s.t. for all $t \geq \gamma$, $\hat{Q}_{\gamma, \eta}(t)$ is at most

$$\eta + \max \left(1 + \phi(\gamma, t - \gamma, \eta), \phi(0, \gamma, n) + A^\gamma(t - \gamma) - \sum_{i=1}^{\eta} N_i^\gamma(t - \gamma) \right) + \sum_{i=\eta+1}^{n} I(V_i^1(\gamma) > t - \gamma).$$

We note that in the above, the term $\phi(0, \gamma, n) + A^\gamma(t - \gamma) - \sum_{i=1}^{\eta} N_i^\gamma(t - \gamma)$ dominates the associated maximum exactly when no artificial arrivals are added on $[\gamma, t]$, in which case $\phi(0, \gamma, n)$ equals the
number of jobs which were waiting in queue at time $\gamma$. Also, the term $\sum_{i=\eta+1}^{n} I(V_i^1(\gamma) > t - \gamma)$ represents the number of jobs which were on “broken servers” $\eta+1, \ldots, n$ at time $\gamma$, that have not completed service by time $t$. Using this bounding system, we can get meaningful bounds for the s.s.p.d. and number of idle servers even when we require using values of $\eta$ which drive the system into instability, as it only operates with this reduced number of servers for a limited time (i.e. on $[\gamma, t]$).

We now formally state our upper bound result, which is informally captured by (2). In our formal upper bound result, several renewal-theoretic quantities appearing in (2) are simplified, e.g. by exploiting the stationarity of equilibrium renewal processes. Then our main upper bound result is as follows.

**Theorem 4.** For all $t, x \geq 0$, $P(Q(t) > x)$ is at most

\[
\inf_{\delta \in [0,t]} \inf_{\eta \in [0,n]} \mathbb{P} \left( \max_{0 \leq \eta \leq \delta} \left( A(s) - \sum_{i=1}^{\eta} N_i(s) \right), \sup_{\delta \leq s \leq t} \left( A(s) - \sum_{i=1}^{\eta} N_i(s) \right) \right) + \sum_{i=\eta+1}^{n} \mathbb{I}(N_i(\delta) = 0) > x - \eta \right).
\]

If in addition $Q(t)$ converges weakly to a steady-state distribution $Q(\infty)$ as $t \to \infty$, then for all $x > 0$, $P(Q(\infty) > x)$ is at most

\[
\inf_{\delta \geq 0} \inf_{\eta \in [0,n]} \mathbb{P} \left( \max_{0 \leq \eta \leq \delta} \left( A(t) - \sum_{i=1}^{\eta} N_i(t) \right), \sup_{t \geq \delta} \left( A(t) - \sum_{i=1}^{\eta} N_i(t) \right) + \sum_{i=\eta+1}^{n} \mathbb{I}(N_i(\delta) = 0) > x - \eta \right).
\]

We note that the upper-bound results of Gamarnik and Goldberg [21] can be recovered as a special case of Theorem 4, modulo an additional “+ 1” which appears in the statement of Theorem 4, by setting $\delta = 0$ and $\eta = n$.

We now prove Theorem 4, and begin by defining a third modified system $\tilde{Q}_{v, \eta}^u$, where the extra $v$ parameter will allow us to consider more general initial conditions, and thus treat all necessary cases using a single notation. Informally, $\tilde{Q}_{v, \eta}^u$ has initial conditions “shifted in time” by $v$, i.e. the remaining processing time of the initial job on server $i$ will be $V_i^1(v)$ instead of $V_i^1$, and there are $\phi(0,v,n)$ jobs initially waiting in queue. Furthermore, $\tilde{Q}_{v, \eta}^u$ will be a modified $\eta$-server system, s.t. on $[0, \gamma]$, servers $1, \ldots, \eta$ are kept busy at all times by adding an artificial arrival whenever one of these servers would otherwise have gone idle. Then, on $[\gamma, t]$, the arrival process to $\tilde{Q}_{v, \eta}^u$ is “unmodified”, i.e. the same arrival process seen by $Q$ on $[\gamma, t]$. Thus, referring to the systems introduced previously, $Q_{v, \eta}^u(t)$ coincides with $\tilde{Q}_{0,n}^u(t)$, $Q(t)$ coincides with $\tilde{Q}_{0,n}^u(t)$, $Q_{\eta}(t)$ coincides with $\tilde{Q}_{1,n}^u(t)$, and $\tilde{Q}_{v, \eta}(t)$ coincides with $\tilde{Q}_{1,v+n}(t - \gamma) + \sum_{i=\eta+1}^{n} I(V_i^1(\gamma) > t - \gamma)$, which follows from the interpretation of the relevant systems as Markov chains on an appropriate state-space.

To formally define $\tilde{Q}_{v, \eta}^u$, we begin by defining two auxiliary processes $\tilde{A}_{v, \eta}^u(t)$ and $\tilde{Q}_{v, \eta}^u(t)$, where $\tilde{A}_{v, \eta}^u(t)$ will become the arrival process to $\tilde{Q}_{v, \eta}^u$, and we will later prove that $\tilde{Q}_{v, \eta}^u(t)$ equals the number in system in $\tilde{Q}_{\gamma, \eta}^u$ at time $t$. Whenever there is no ambiguity, we use the notations $\tau_k$, $\tilde{Q}$, $\tilde{A}$, $\tilde{Q}(t)$, and $\tilde{Q}(\tau_k)$ as shorthand for $\tau_{v,k}$, $\tilde{Q}_{v, \eta}^u(\tau_k)$, $\tilde{A}_{v, \eta}^u$, and $\tilde{Q}_{v, \eta}^u(t)$ respectively.

We now define the processes $\tilde{A}(t)$ and $\tilde{Q}(t)$ on $[0, \gamma]$ inductively over $\{\tau_k, k \geq 0\}$. Let $\tilde{A}(\tau_0) \triangleq 0$, $\tilde{Q}(\tau_0) \triangleq \eta + \phi(0,v,n)$. Now suppose that for some $k \geq 0$, we have defined $\tilde{A}(t)$ and $\tilde{Q}(t)$ for all $t \leq \tau_k$, and $\tau_{k+1} \leq \gamma$. We now define these processes for $t \in (\tau_k, \tau_{k+1}]$. For $t \in (\tau_k, \tau_{k+1})$, let $\tilde{A}(t) \triangleq \tilde{A}(\tau_k)$,
and $\hat{Q}(t) \triangleq \hat{Q}(\tau_k)$. Note that $dA^\nu(\tau_{k+1}) + \sum_{i=1}^\eta dN^\nu_i(\tau_{k+1}) = 1$, since $R(S)$ and $R(A)$ are continuous r.v.s, $P(A = 0) = P(S = 0) = 0$, and $A(t), \{N_i(t), i = 1, \ldots, \eta\}$ are mutually independent and have stationary increments. We define

$$
\hat{A}(\tau_{k+1}) \triangleq \begin{cases} 
\hat{A}(\tau_k) + 1 & \text{if } dA^\nu(\tau_{k+1}) = 1; \\
\hat{A}(\tau_k) + 1 & \text{if } \sum_{i=1}^\eta dN^\nu_i(\tau_{k+1}) = 1 \text{ and } \hat{Q}(\tau_k) \leq \eta; \\
\hat{A}(\tau_k) & \text{otherwise (i.e. } \sum_{i=1}^\eta dN^\nu_i(\tau_{k+1}) = 1 \text{ and } \hat{Q}(\tau_k) > \eta). 
\end{cases}
$$

Similarly, we define

$$
\hat{Q}(\tau_{k+1}) \triangleq \begin{cases} 
\hat{Q}(\tau_k) + 1 & \text{if } dA^\nu(\tau_{k+1}) = 1; \\
\hat{Q}(\tau_k) & \text{if } \sum_{i=1}^\eta dN^\nu_i(\tau_{k+1}) = 1 \text{ and } \hat{Q}(\tau_k) \leq \eta; \\
\hat{Q}(\tau_k) - 1 & \text{otherwise (i.e. } \sum_{i=1}^\eta dN^\nu_i(\tau_{k+1}) = 1 \text{ and } \hat{Q}(\tau_k) > \eta). 
\end{cases}
$$

Also, for $k = A^\nu(\gamma) + \sum_{i=1}^\gamma N^\nu_i(\gamma)$, namely the largest index s.t. $\tau_k \leq \gamma$, we let $\hat{A}(t) \triangleq \hat{A}(\tau_k)$ for $t \in (\tau_k, \gamma]$, and $\hat{Q}(t) \triangleq \hat{Q}(\tau_k)$ for $t \in (\tau_k, \gamma]$. Combining the above completes our inductive definition of $\hat{A}(t)$ and $\hat{Q}(t)$ on $[0, \gamma]$, since $\lim_{k \to \infty} \tau_k = \infty$.

We now define $\hat{A}(t)$ on $(\gamma, \infty)$. For all $t > \gamma$, $d\hat{A}(t) = dA^\nu(t)$, namely the events in the two processes coincide. It follows from our construction that both $\hat{A}(t)$ and $\hat{Q}(t)$ are well-defined and r.c.l.l. on $[0, \gamma]$, and $\hat{A}(t)$ is well-defined and r.c.l.l. on $[0, \infty)$.

We now construct the FCFS $G/G/\eta$ queue $\hat{Q}$ using the auxiliary process $\hat{A}(t)$, and define $\hat{Q}(t)$ on $[\gamma, \infty)$. $\hat{Q}$ is defined to be the FCFS $G/G/\eta$ queue with arrival process $\hat{A}(t)$ and processing time distribution $S$, where the $j$th job assigned to server $i$ (after time 0) is assigned processing time $V_i^j(\nu)$ for $j \geq 1, i = 1, \ldots, \eta$, and jobs are always assigned to the available server of least index. The initial conditions for $\hat{Q}$ are s.t. for $i = 1, \ldots, \eta$, there is a single job initially being processed on server $i$ with initial processing time $V_i^1(\nu)$, and there are $\phi(0, \nu, n)$ jobs waiting in queue. For $t \geq \gamma$, we define $\hat{Q}(t)$ to be the number in system in $\hat{Q}$ at time $t$. Note that on $[\gamma, \infty)$, $\hat{Q}$ operates like a “normal” FCFS $GI/GI/\eta$ queue.

We now analyze $\hat{Q}$ on $[0, \gamma]$. The following lemma is essentially identical to Lemma 1 of Gamarnik and Goldberg [21]. The proof follows from a straightforward induction nearly identically to the proof of Lemma 1 in Gamarnik and Goldberg [21], and we refer the reader to Gamarnik and Goldberg [21] for details.

**Lemma 1.** For $i = 1, \ldots, \eta$, exactly one job departs from server $i$ at each time $t \in \{\sum_{j=1}^j V_i^j(\nu), j \geq 1\} \cap [0, \gamma]$, and there are no other departures from server $i$ on $[0, \gamma]$. Also, no server ever idles in $\hat{Q}$ on $[0, \gamma]$, $\hat{Q}(t)$ equals the number in system in $\hat{Q}$ at time $t$ for all $t \leq \gamma$, and for all $k$ s.t. $\tau_k \leq \gamma$,

$$
\hat{Q}(\tau_k) - \eta = \max \left(0, \hat{Q}(\tau_{k-1}) - \eta + dA^\nu(\tau_k) - \sum_{i=1}^\eta dN_i^\nu(\tau_k)\right).
$$

Note that it follows from Lemma 1 and our definition of $\hat{Q}(t)$ on $[\gamma, \infty)$ that $\hat{Q}(t)$ is r.c.l.l. on $[0, \infty)$, and

**Corollary 1.** $\hat{Q}(t)$ equals the number in system in $\hat{Q}$ at time $t$ for all $t \geq 0$.

We now ‘unfold’ recursion (5) to derive a simple one-dimensional random walk representation for $\hat{Q}(t)$, $t \leq \gamma$. We note that the relationship between recursions such as (5) and the suprema of associated one-dimensional random walks is well-known (see Borovkov [5], Chang et al. [10], Gamarnik and Goldberg [21]), and can also be formalized by studying the appropriate Skorokhod problem - we refer the reader to Skorokhod [53] for details.
It follows from (5) and a straightforward induction on \( \{ \tau_k, k \geq 0 \} \) that for all \( k \) s.t. \( \tau_k \leq \gamma \),
\[
\hat{Q}(\tau_k) - \eta = \max\left( \max_{j \in [0,k-1]} \left( A^v(\tau_{k-j}, \tau_k) - \sum_{i=1}^\eta N^v_i(\tau_{k-j}, \tau_k) \right), \hat{Q}(0) - \eta + A^v(\tau_k) - \sum_{i=1}^\eta N^v_i(\tau_k) \right).
\]
As all jumps in \( \hat{Q}(t) \) on \( [0, \gamma] \) occur at times \( t \in \{ \tau_k, k \geq 1 \} \), and are of size 1,
\[
\max_{j \in [0,k-1]} \left( A^v(\tau_{k-j}, \tau_k) - \sum_{i=1}^\eta N^v_i(\tau_{k-j}, \tau_k) \right) \leq 1 + \max_{j \in [0,k]} \left( A^v(\tau_{k-j}, \tau_k) - \sum_{i=1}^\eta N^v_i(\tau_{k-j}, \tau_k) \right).
\]
It follows that

**Corollary 2.** If \( \hat{Q}(0) = \eta \), then \( \hat{Q}(\gamma) - \eta = \phi(v, \gamma, \eta) \). In general,
\[
\hat{Q}(\gamma) - \eta \leq \max \left( 1 + \phi(v, \gamma, \eta), \phi(0, v, n) + A^v(\gamma) - \sum_{i=1}^\eta N^v_i(\gamma) \right).
\]

Before proceeding, it will be useful to prove a general comparison result for \( G/G/n \) queues, similar to Lemma 2 of Gamarnik and Goldberg [21]. The key difference is that here we allow for both general initial conditions and differing numbers of servers. For an event \( \{ E \} \), let \( I(\{ E \}) \) denote the indicator function of \( \{ E \} \). Then

**Lemma 2.** Let \( Q^1 \) denote a FCFS \( G/G/n^1 \) queue, and \( Q^2 \) denote a FCFS \( G/G/n^2 \) queue, both with finite, strictly positive inter-arrival and processing times, s.t. \( n^1 \geq n^2 \). Let \( Q^1(t) \) denote the number in system at time \( t \) in \( Q^i \), and \( L^i = Q^i(0) - n^i \), \( i \in \{ 1, 2 \} \). For \( k \in \{ 1, \ldots, L^2 \} \), let \( T^1_k \) equal zero, and \( S^i_k \) denote the initial processing time of the \( k \)th job initially waiting in queue in \( Q^i \), \( i \in \{ 1, 2 \} \). For \( k > L^2 \), let \( T^1_k \) denote the arrival time of the \( (k-L^2) \)th arrival (after time 0) to \( Q^2 \), and \( S^i_k \) the processing time assigned to that job, \( i \in \{ 1, 2 \} \). Further suppose that

(i) \( L^1 = L^2 \geq 0 \), and we denote this common value by \( L \). Also, \( S^1_k = S^2_k \) for \( k \in \{ 1, \ldots, L \} \). That is, the \( k \)th job initially waiting in queue in \( Q^2 \) is assigned the same processing time as the \( k \)th job initially waiting in queue in \( Q^1 \), \( k = 1, \ldots, L \). In addition, we let \( W_i \) denote the initial processing time of the job initially being processed on server \( i \) in \( Q^1 \), \( i \in \{ 1, \ldots, n^1 \} \).

(ii) For each job \( J \) initially being processed in \( Q^1 \) on a server whose index belongs to the set \( \{ 1, \ldots, n^2 \} \), there is a distinct corresponding job \( J' \) initially being processed in \( Q^2 \), s.t. the initial processing time of \( J \) in \( Q^1 \) equals the initial processing time of \( J' \) in \( Q^2 \).

(iii) \( \{ T^1_k, k \geq 1 \} \) is a subsequence of \( \{ T^2_k, k \geq 1 \} \).

(iv) For all \( k > L, \) the job that arrives to \( Q^2 \) at time \( T^2_k \) is assigned processing time \( S^1_k \), the same processing time assigned to the job which arrives to \( Q^1 \) at that time.

Then for all \( t \geq 0 \),
\[
Q^1(t) \leq Q^2(t) + \sum_{i=n^2+1}^{n^1} I(W_i > t).
\]

**Proof** Let \( Z^1_i(t) \) denote the number of jobs initially being processed in \( Q^1 \), on servers with index \( i \leq n^2 \), which are still in \( Q^1 \) at time \( t \). Let \( Z^2_i(t) \) denote the number of jobs initially being processed in \( Q^1 \), on servers with index \( i > n^2 \), which are still in \( Q^1 \) at time \( t \). Let \( Z^2(t) \) denote the number of jobs initially being processed in \( Q^2 \), which are still in \( Q^2 \) at time \( t \). Note that by (ii), \( Z^1_i(t) = Z^2(t) \) for all \( t \geq 0 \), and we denote this common value by \( Z(t) \). Also, for all \( t \geq 0 \),
\[
Z^2(t) = \sum_{i=n^2+1}^{n^1} I(W_i > t).
\]
For $k \in \{1, \ldots, \mathcal{L}\}$, let $D_k^i$ denote the time at which the $k$th job initially waiting in queue in $Q^i$ departs from $Q^i$, $i \in \{1, 2\}$. For $k > \mathcal{L}$, let $D_k^i$ denote the time at which the job that arrives to $Q^i$ at time $T_k^i$ departs from $Q^i$, $i \in \{1, 2\}$. Observe that for all $k \geq 1$

\[
D_k^1 = \inf \{ t : t \geq T_k^1, Z(t) + Z_k^1(t) + \sum_{j=1}^{k-1} I(D_j^1 > t) \leq n^1 - 1\} + S_k^1
\]

\[
= \inf \{ t : t \geq T_k^1, Z(t) + \sum_{j=1}^{k-1} I(D_j^1 > t) \leq n^1 - Z_k^1(t) - 1\} + S_k^1
\]

\[
\leq \inf \{ t : t \geq T_k^1, Z(t) + \sum_{j=1}^{k-1} I(D_j^1 > t) \leq n^2 - 1\} + S_k^1.
\]

(6)

Also,

\[
D_k^2 \geq \inf \{ t : t \geq T_k^1, Z(t) + \sum_{j=1}^{k-1} I(D_j^2 > t) \leq n^2 - 1\} + S_k^1,
\]

(7)

where the inequality in (7) arises since the job that arrives to $Q^2$ at time $T_k^1$ may have to wait for additional jobs, which arrive at a time belonging to $\{T_k^1, k \geq 1\} \setminus \{T_k^1, k \geq 1\}$.

We now prove by induction that for $k \geq 1$, $D_k^2 \geq D_k^1$, from which the lemma follows. For the base case $k = 1$, note that $D_1^1 \leq \inf \{ t : t \geq T_1^1, Z(t) \leq n^2 - 1\} + S_1^1$ by (6), while $D_1^2 \geq \inf \{ t : t \geq T_1^1, Z(t) \leq n^2 - 1\} + S_1^1$ by (7).

Now assume the induction is true for all $j \leq k$. Then for all $t \geq 0$, $\sum_{j=1}^{k} I(D_j^2 > t) \geq \sum_{j=1}^{k} I(D_j^1 > t)$. Thus

\[
\inf \{ t : t \geq T_{k+1}^1, Z(t) + \sum_{j=1}^{k} I(D_j^1 > t) \leq n^2 - 1\} + S_{k+1}^1
\]

\[
\leq \inf \{ t : t \geq T_{k+1}^1, Z(t) + \sum_{j=1}^{k} I(D_j^2 > t) \leq n^2 - 1\} + S_{k+1}^1.
\]

It then follows from (6) and (7) that $D_{k+1}^1 \leq D_{k+1}^2$, completing the induction. \(\square\)

We now complete the proof of Theorem 4.

**Proof** [Proof of Theorem 4] Let us fix some $\gamma, \tau \in \mathbb{R}^+$ s.t. $\gamma \leq \tau$, and $\eta \in \mathbb{Z}^+$ s.t. $\eta \leq n$. We begin by constructing $Q$ and $\tilde{Q}_{\gamma,n}^0$ on the same probability space. Note that $Q$ and $\tilde{Q}_{\gamma,n}^0$ have the same initial conditions. Namely, for $i = 1, \ldots, n$, there is a single job initially being processed on server $i$ with initial processing time $V_i^1$, there are 0 jobs waiting in queue, and the time until the first arrival is $U^1$. We let $A(t)$ be the arrival process to $Q$. Let $\{t_k, k \geq 1\}$ denote the ordered sequence of event times in $A(t)$. It follows from the construction of $A_{\gamma,n}^0(t)$ that $\{t_k, k \geq 1\}$ is a subsequence of the set of event times in $A_{\gamma,n}^0(t)$. We let the processing time assigned to the arrival to $\tilde{Q}_{\gamma,n}^0$ at time $t_k$ equal the processing time assigned to the arrival to $Q$ at time $t_k$, $k \geq 1$. It follows that $Q$ and $\tilde{Q}_{\gamma,n}^0$ satisfy the conditions of Lemma 2, and it follows from Corollary 1 that for all $x \geq 0$, $\mathbb{P}(Q(\tau) > x) \leq \mathbb{P}(\tilde{Q}_{\gamma,n}^0(\tau) > x)$.

(8)

It follows from Lemma 1 that at time $\gamma$, server $i$ of $\tilde{Q}_{\gamma,n}^0$ is processing a job with remaining processing time (at time $\gamma$) $V_i^1(\gamma)$, $i = 1, \ldots, n$. It follows from Corollary 2 that at time $\gamma$, there are $\phi(0, \gamma, n)$ jobs waiting in queue in $\tilde{Q}_{\gamma,n}^0$. Also, the remaining time (at time $\gamma$) until the next arrival to $\tilde{Q}_{\gamma,n}^0$ is $U^1(\gamma)$. Thus by construction, the state of $\tilde{Q}_{\gamma,n}^0$ at time $\gamma$ (viewed as a Markov
chain on an appropriate state space) is identical to the state of $\mathcal{Q}_n^\gamma$ at time 0. It then follows from our construction, the Markov chain interpretation of the $GI/GI/n$ queue (see Asmussen [2]), and Corollary 1 that we may construct $\mathcal{Q}_n^0$ and $\mathcal{Q}_n^\gamma$ on the same probability space s.t. $\mathcal{Q}_n^0(\gamma + s) = \mathcal{Q}_n^\gamma(s)$ for all $s \geq 0$. Thus for all $x \geq 0$,

$$\mathbb{P}(\mathcal{Q}_n^0(\tau) > x) = \mathbb{P}(\mathcal{Q}_n^\gamma(\tau - \gamma) > x).$$

(9)

We now construct $\mathcal{Q}_n^\gamma$ and $\mathcal{Q}_n^\gamma$ on the same probability space $\Omega$. Suppose w.l.o.g. that $A(t)$, \{N_i(t), i = 1, \ldots, n\}, and all associated auxiliary r.v.s (e.g. $A^\gamma(t)$, \{N_i^\gamma(t), i = 1, \ldots, n\}, \{V_i^\gamma(\gamma), i = 1, \ldots, n, j \geq 1\}, $\phi(0, \gamma, n)$ have been constructed on $\Omega$. Note that for $i \leq \eta$, the initial processing time of the job initially being processed on server $i$ is $V_i^\gamma(\gamma)$ in both $\mathcal{Q}_n^\gamma$ and $\mathcal{Q}_n^\gamma$. Also, in both systems there are $\phi(0, \gamma, n)$ jobs initially waiting in queue, and we assign corresponding initial jobs waiting in queue the same processing time. We let $A^\gamma(t)$ be the arrival process to both systems, and assign the same processing time to corresponding arrivals. It follows that on $\Omega$, $\mathcal{Q}_n^\gamma$ and $\mathcal{Q}_n^\gamma$ satisfy the conditions of Lemma 2, and

$$\mathcal{Q}_n^\gamma(\tau - \gamma) \leq \mathcal{Q}_n^\gamma(\tau - \gamma) + \sum_{i=\eta+1}^{n} I(V_i^\gamma(\gamma) > \tau - \gamma).$$

(10)

Note that on $\Omega$, all inter-arrival times for jobs arriving to $\mathcal{Q}_n^\gamma$ (except the first), and all processing times assigned to jobs not initially being processed in $\mathcal{Q}_n^\gamma$, are independent of $\{V_i^\gamma(\gamma), i = 1, \ldots, n\}$, and thus independent of $\sum_{i=\eta+1}^{n} I(V_i^\gamma(\gamma) > \tau - \gamma)$. Also note that although our construction ensures that the assignment of processing times in both $\mathcal{Q}_n^\gamma$ and $\mathcal{Q}_n^\gamma$ coincide, we have not yet specified any particular construction for these processing times on $\Omega$.

We now construct $\tilde{\mathcal{Q}}_{\tau - \gamma, \eta}$ on the same probability space $\Omega$, and simultaneously give an explicit construction for the assignment of processing times to jobs for $\mathcal{Q}_n^\gamma$ on $\Omega$. Let us construct $\tilde{\mathcal{Q}}_{\tau - \gamma, \eta}$ by constructing the auxiliary processes $\tilde{A}_{\tau - \gamma, \eta}(t)$ and $\tilde{Q}_{\tau - \gamma, \eta}(t)$, exactly as in (3) - (4), using the same primitive processes (already constructed on $\Omega$) $A(t)$, \{N_i(t), i = 1, \ldots, n\}, and the associated auxiliary r.v.s. Note that by construction, $\mathcal{Q}_n^\gamma$ and $\tilde{\mathcal{Q}}_{\tau - \gamma, \eta}$ have the same initial conditions, and thus the same number $\phi(0, \gamma, n)$ of jobs initially waiting in queue. We assign corresponding initial jobs waiting in queue the same processing time, as dictated by the assignment of processing times to jobs in $\tilde{\mathcal{Q}}_{\tau - \gamma, \eta}$. Namely, if the $k$th job initially waiting in queue in $\tilde{\mathcal{Q}}_{\tau - \gamma, \eta}$ is the $j$th job assigned to server $i$ in $\tilde{\mathcal{Q}}_{\tau - \gamma, \eta}$, it receives processing time $V_i^{\gamma+1}(\gamma)$. Recall that $A^\gamma(t)$ is the arrival process to $\mathcal{Q}_n^\gamma$. It follows from the construction of $\tilde{A}_{\tau - \gamma, \eta}(t)$ that the sequence of arrival times to $\mathcal{Q}_n^\gamma$ is a subsequence of the set of arrival times to $\tilde{\mathcal{Q}}_{\tau - \gamma, \eta}$. Also, in both systems there are $\phi(0, \gamma, n)$ jobs initially waiting in queue, and we assign corresponding initial jobs waiting in queue the same processing time. We let $A^\gamma(t)$ be the arrival process to both systems, and assign the same processing time to corresponding arrivals. Namely, if the corresponding arrival in $\tilde{\mathcal{Q}}_{\tau - \gamma, \eta}$ is the $j$th job assigned to server $i$ in $\tilde{\mathcal{Q}}_{\tau - \gamma, \eta}$, it receives processing time $V_i^{\gamma+1}(\gamma)$. It follows that on the probability space $\Omega$, $\mathcal{Q}_n^\gamma$ and $\tilde{\mathcal{Q}}_{\tau - \gamma, \eta}$ satisfy the conditions of Lemma 2, and it thus follows from (10) and Corollary 1 that

$$\mathcal{Q}_n^\gamma(\tau - \gamma) \leq \tilde{\mathcal{Q}}_{\tau - \gamma, \eta}(\tau - \gamma) + \sum_{i=\eta+1}^{n} I(V_i^\gamma(\gamma) > \tau - \gamma).$$

(11)

Combining (8) - (11) with Corollary 2, and observing that $\gamma, \tau, \eta$ were general, we find that for all $t, x \geq 0$, $\mathbb{P}(Q(t) > x)$ is at most

$$\inf_{\gamma \in [0,t]} \inf_{\eta \in [0,n]} \mathbb{P} \left( \max \left( 1 + \phi(\gamma, t - \gamma, \eta), \phi(0, \gamma, n) + A^\gamma(t - \gamma) - \sum_{i=\eta+1}^{n} N_i^\gamma(t - \gamma) \right) \right)
\sum_{i=\eta+1}^{n} I(V_i^\gamma(\gamma) > t - \gamma) > x - \eta. $$

(12)
From definitions,
\[
\phi(\gamma, t - \gamma, \eta) = \sup_{0 \leq s \leq t - \gamma} (A(t - s, t) - \sum_{i=1}^{\eta} N_i(t - s, t));
\]
\[
\phi(0, \gamma, n) = \sup_{0 \leq s \leq \gamma} (A(\gamma - s, \gamma) - \sum_{i=1}^{n} N_i(\gamma - s, \gamma));
\]
\[
A^\gamma(t - \gamma) - \sum_{i=1}^{\gamma} N_i^\gamma(t - \gamma) = A(\gamma, t) - \sum_{i=1}^{\gamma} N_i(\gamma, t);
\]
\[
\sum_{i=\eta+1}^{n} I(V_i^1(\gamma) > t - \gamma) = \sum_{i=\eta+1}^{n} I(N_i(\gamma, t) = 0).
\]

It follows from elementary renewal theory (see Cox [11]) that the joint distribution of
\[
\phi(\gamma, t - \gamma, \eta), \phi(0, \gamma, n), A^\gamma(t - \gamma) - \sum_{i=1}^{\eta} N_i^\gamma(t - \gamma), \sum_{i=\eta+1}^{n} I(V_i^1(\gamma) > t - \gamma)
\]
is the same as that of
\[
\sup_{0 \leq s \leq t - \gamma} (A(s) - \sum_{i=1}^{\eta} N_i(s)), \sup_{0 \leq s \leq \gamma} (A^{t-\gamma}(s) - \sum_{i=1}^{n} N_i^{t-\gamma}(s)),
\]
\[
A(t - \gamma) - \sum_{i=1}^{\gamma} N_i(t - \gamma), \sum_{i=\eta+1}^{n} I(N_i(t - \gamma) = 0).
\]

Combining the above with (12), letting \( \delta \Delta = t - \gamma \), and simplifying the resulting expressions completes the proof of the first part of the theorem.

We now prove the corresponding steady-state result. Note that for all \( \delta \leq t \),
\[
\sup_{\delta \leq s \leq t} (A(s) - \sum_{i=1}^{n} N_i(s)) \leq \sup_{s \geq \delta} (A(s) - \sum_{i=1}^{n} N_i(s)).
\]

It follows that for any fixed values of \( \eta \) and \( x \), \( \mathbb{P}(Q(t) > x) \) is at most
\[
\inf_{\delta \in [0,t]} \mathbb{P} \left( \max \left( 1 + \sup_{0 \leq s \leq \delta} (A(s) - \sum_{i=1}^{\eta} N_i(s)), \sup_{s \geq \delta} (A(s) - \sum_{i=1}^{n} N_i(s)) + \sum_{i=\eta+1}^{n} N_i(\delta) \right) + \sum_{i=\eta+1}^{n} I(N_i(\delta) = 0) > x - \eta \right) \tag{13}
\]

Since (13) is monotone decreasing in \( t \), the corresponding limit exists as \( t \to \infty \), completing the proof. \( \square \)

4. Asymptotic Bound in the Halfin-Whitt Regime

In this section, we use Theorem 4 to bound the FCFS \( GI/GI/n \) queue in the H-W regime, by proving an asymptotic analogue of Theorem 4. Suppose that the H-W and \( T_0 \) assumptions hold. For a r.v. \( X \), let \( V[X] \) denote the variance of \( X \). For r.v.s \( X, Y \), let \( V[X, Y] \triangleq \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \) denote the covariance of \( X \) and \( Y \). Recall that a Gaussian process on \( \mathbb{R} \) is a stochastic process \( Z(t)_{t \geq 0} \) s.t. for any finite set of times \( t_1, \ldots, t_k \), the vector \( (Z(t_1), \ldots, Z(t_k)) \) has a Gaussian distribution.
A Gaussian process $Z(t)$ is known to be uniquely determined by its mean function $\mathbb{E}[Z(t)]$ and covariance function $V[Z(s),Z(t)]$, and refer the reader to Doob [18], Ibragimov and Rozanov [31], Adler [1], Marcus and Rosen [43], and the references therein for details on existence, continuity, etc. It is proven by Whitt [58] Theorem 2 that there exists a continuous Gaussian process $D(t)$ s.t. $\mathbb{E}[D(t)]=0, V[D(s),D(t)]=\mathbb{E}[(N_1(s)-\mu s)(N_1(t)-\mu t)]$ for all $s,t \geq 0$. Let $A(t)$ denote the continuous Gaussian process s.t. $\mathbb{E}[A(t)]=0, V[A(s),A(t)]=\mu c^2 \min(s,t)$, namely $A(t)$ is a driftless Brownian motion. Let $Z(t) \overset{d}{=} A(t) - D(t)$, where $A(t)$ and $D(t)$ are independent. Let us define the event

$$
\mathcal{E}_{B,\infty}^{\delta,\eta}(x) \overset{d}{=} \{ \max \left( \sup_{0 \leq t \leq \delta} (Z(t) + (\eta - B)\mu t), \sup_{t \geq \delta} (Z(t) - B\mu t + \eta \mu \delta) \right) \geq x + \eta \mathbb{P}(B(S) \leq \delta) \}.
$$

Then our main asymptotic upper bound is that

**Theorem 5.** For all $B > 0$ and $x \in \mathbb{R}$,

$$\limsup_{n \to \infty} \mathbb{P}\left( (Q_B^n(\infty) - n)n^{-\frac{1}{2}} \geq x \right) \leq \inf_{\eta \geq 0} \mathbb{P}(\mathcal{E}_{B,\infty}^{\delta,\eta}(x)).$$

Before proving Theorem 5, we first introduce some additional notations. Let $A^n_B(t)$ denote an equilibrium renewal process with renewal distribution $A_{n, B}^{-1}$, independent of $\{N_i(t), i = 1, \ldots, n\}$, $Z^n_B(t) \overset{d}{=} A^n_B(t) - \sum_{i=1}^{n} N_i(t)$, and $Z^n_B(s, t) \overset{d}{=} Z^n_B(t) - Z^n_B(s)$. Also, we define $f_n(x) \overset{d}{=} [n - xn^{\frac{1}{2}}]$, and

$$
\mathcal{E}_{B,n}^{\delta,\eta}(x) \overset{d}{=} \left\{ \max \left( 1 + \sup_{0 \leq t \leq \delta} (Z^n_B(t) + \sum_{i=f_n(\eta)+1}^{n} N_i(t)), \sup_{t \geq \delta} (Z^n_B(t) + \sum_{i=f_n(\eta)+1}^{n} N_i(\delta)) \right) + \sum_{i=f_n(\eta)+1}^{n} I(N_i(\delta) = 0) \geq f_n(-x) - f_n(\eta) \right\}.
$$

Then it follows from Theorem 4 that for all $x \in \mathbb{R}$,

$$\mathbb{P}\left( (Q^n_B(\infty) - n)n^{-\frac{1}{2}} \geq x \right) \leq \inf_{\eta \in [0, n^{\frac{1}{2}}]} \mathbb{P}(\mathcal{E}_{B,n}^{\delta,\eta}(x)).$$

Before completing the proof of Theorem 5, we establish some preliminary weak convergence results to aid in the analysis of $\mathbb{P}(\mathcal{E}_{B,n}^{\delta,\eta}(x))$.

### 4.1. Preliminary weak convergence results

For an excellent review of weak convergence, and the associated spaces (e.g. $D(0,T)$) and topologies/metrics (e.g. uniform, $J_1$), the reader is referred to Whitt [59]. We now review several results of Gamarnik and Goldberg [21], for use in our analysis.

**Theorem 6.** (i) For any fixed $T \in [0, \infty)$, the sequence of processes $\{n^{-\frac{1}{2}}(\sum_{i=1}^{n} N_i(t) - n\mu t)\}_{0 \leq t \leq T}, n \geq 1$ converges weakly to $D(t)_{0 \leq t \leq T}$ in the space $D[0,T]$ under the $J_1$ topology.

(ii) For any fixed $B > 0$ and $T \in [0, \infty)$, the sequence of processes $\{n^{-\frac{1}{2}}Z^n_B(t)_{0 \leq t \leq T}, n \geq 1\}$ converges weakly to $(Z(t) - B\mu t)_{0 \leq t \leq T}$ in the space $D[0,T]$ under the $J_1$ topology.

(iii) For any fixed $B, x > 0$, $\limsup_{T \to \infty} \limsup_{n \to \infty} \mathbb{P}(n^{-\frac{1}{2}} \sup_{t \geq T} Z^n_B(t) \geq x) = 0$. It follows that the sequence of r.v.s $\{n^{-\frac{1}{2}} \sup_{t \geq T} Z^n_B(t), n \geq 1\}$ is tight.

(iv) $\lim_{t \to \infty} \mathbb{E}[\left( t^{-\frac{1}{2}} Z(t) \right)^2] = \mu(c_A^2 + c_S^2)$. Also, the r.v. $\sup_{t \geq 0} (Z(t) - B\mu t)$ is a.s. finite.
Proof (i) follows from Gamarnik and Goldberg [21] Theorem 7. (ii) follows from Gamarnik and Goldberg [21] Lemma 6. (iii) follows from Gamarnik and Goldberg [21] Lemma 7. (iv) follows from Gamarnik and Goldberg [21] Equation 17, and Corollary 3. □

We now state several additional weak convergence results, also for use in our analysis.

**Corollary 3.** (i) For any fixed \( B > 0 \), \( \delta \geq 0 \), and \( \epsilon > 0 \), there exists \( T \in (\delta, \infty) \) s.t.

\[
\limsup_{n \to \infty} \mathbb{P}\left( \sup_{t \geq \delta} Z_B^n(t) > \sup_{t \in [\delta, T]} Z_B^0(t) \right) < \epsilon.
\]

(ii) \( \lim_{n \to \infty} n^{-\frac{1}{2}} (n - f_n(\eta)) = \eta \), and \( \lim_{n \to \infty} n^{-\frac{1}{2}} (f_n(-x) - f_n(\eta)) = x + \eta \).

(iii) The sequence of r.v.s \( \left\{ n^{-\frac{1}{2}} \sum_{i=f_n(\eta)+1}^{n} N_i(\delta), n \geq 1 \right\} \) converges weakly to the constant \( \eta \mu \delta \).

(iv) The sequence of r.v.s \( \left\{ n^{-\frac{1}{2}} \sum_{i=f_n(\eta)+1}^{n} I(N_i(\delta) = 0), n \geq 1 \right\} \) converges weakly to the constant \( \eta \mathbb{P}(R(S) > \delta) \).

(v) For all \( \eta \geq 0 \) and \( T \in [0, \infty) \), the sequence of processes \( \left\{ n^{-\frac{1}{2}} \left( \sum_{i=f_n(\eta)+1}^{n} N_i(t) - (n - f_n(\eta)) \mu \right) \right\}_{0 \leq t \leq T}, n \geq 1 \) converges weakly to the constant process \( 0 \) in the space \( D[0,T] \) under the \( J_1 \) topology.

Proof (i) follows from Theorem 6.(ii) - (iv). (ii) follows from a straightforward limit. (iii) - (iv) follow from the strong law of large numbers. (v) follows from Theorem 6.(i). □

4.2. Proof of Theorem 5

We now complete the proof of Theorem 5.

Proof [Proof of Theorem 5] Note that it suffices to demonstrate that for each fixed \( B > 0 \) and \( \delta, \eta \geq 0 \), \( \limsup_{n \to \infty} \mathbb{P}\left( \mathcal{E}_{B,n}^\delta,\eta(x) \right) \leq \mathbb{P}\left( \mathcal{E}_{B,\infty}^\delta,\eta(x) \right) \). Let us fix some \( B > 0 \) and \( \delta, \eta \geq 0 \). Applying Corollary 3.(i) and a union bound, and multiplying through by \( n^{-\frac{1}{2}} \), it thus suffices to demonstrate that

\[
\limsup_{T \to \infty} \limsup_{n \to \infty} \mathbb{P}\left( n^{-\frac{1}{2}} \max \left( 1 + \sup_{0 \leq t \leq \delta} \left( Z_B^n(t) + \sum_{i=f_n(\eta)+1}^{n} N_i(t) \right), \sup_{\delta \leq t \leq T} Z_B^n(t) + \sum_{i=f_n(\eta)+1}^{n} N_i(\delta) \right) + n^{-\frac{1}{2}} \sum_{i=f_n(\eta)+1}^{n} I(N_i(\delta) = 0) - n^{-\frac{1}{2}} \left( f_n(-x) - f_n(\eta) \right) \geq 0 \right) \leq \mathbb{P}\left( \mathcal{E}_{B,\infty}^\delta,\eta(x) \right).
\]

(14)

For any fixed \( T \), it follows from Theorem 6.(ii), Corollary 3.(iii)-(v), and the continuity of the supremum map in the space \( D[0,T] \) under the \( J_1 \) topology (see Whitt [59] Theorem 13.4.1) that

\[
\left\{ n^{-\frac{1}{2}} \max \left( 1 + \sup_{0 \leq t \leq \delta} \left( Z_B^n(t) + \sum_{i=f_n(\eta)+1}^{n} N_i(t) \right), \sup_{\delta \leq t \leq T} Z_B^n(t) + \sum_{i=f_n(\eta)+1}^{n} N_i(\delta) \right), n \geq 1 \right\}
\]

converges weakly to \( \max \left( \sup_{0 \leq t \leq \delta} (Z(t) + (\eta - B) \mu t), \sup_{\delta \leq t \leq T} (Z(t) - B \mu t) + \eta \mu \delta \right) \). Inequality (14), and the desired theorem, then follows from Corollary 3.(ii) - (iv), and the Portmanteau Theorem (see Billingsley [3]). □

5. Proof of Bound for Probability of Delay as \( B \to \infty \)

In this section we complete the proof of Theorem 1. We begin by proving a modified variant of Theorem 5, which has the interpretation of setting \( \delta = \infty, \eta = \frac{B}{2} \) in Theorem 5.
Corollary 4. For all $B > 0$ and $x \in \mathbb{R}$, \( \limsup_{n \to \infty} \mathbb{P}\left( (Q^n_B(\infty) - n^{-\frac{1}{2}}) \geq x \right) \) is at most
\[
\mathbb{P}\left( \sup_{t \geq 0} (\mathcal{Z}(t) - B \mu t) \geq x + \frac{B}{2} \right).
\]

Proof It follows from setting \( \eta = \frac{B}{2} \) in Theorem 5, combined with the monotonicity of the supremum operator and a union bound, that for all $B > 0$ and $x \in \mathbb{R}$, \( \limsup_{n \to \infty} \mathbb{P}\left( (Q^n_B(\infty) - n^{-\frac{1}{2}}) > x \right) \) is at most
\[
\inf_{\delta \geq 0} \left( \mathbb{P}\left( \sup_{t \geq 0} (\mathcal{Z}(t) - B \mu t) \geq x + \frac{B}{2} \mathbb{P}(R(S) \leq \delta) \right) + \mathbb{P}\left( \sup_{t \geq \delta} (\mathcal{Z}(t) - B \mu t) \geq -\frac{B}{2} \mu \delta \right) \right).
\]

Let \( \mathcal{Z}(s, t) = \mathcal{Z}(t) - \mathcal{Z}(s) \). Then it follows from the fact that \( \mathcal{Z}(t) \) has stationary increments and a union bound that (16) is at most
\[
\mathbb{P}\left( \mathcal{Z}(\delta) \geq \frac{B}{4} \mu \delta \right) + \mathbb{P}\left( \sup_{t \geq 0} (\mathcal{Z}(t) - B \mu t) \geq \frac{B}{4} \mu \delta \right).
\]

It follows from Theorem 6.(iv) that for any \( \epsilon > 0 \), there exists \( \delta_\epsilon < \infty \) s.t. for all \( \delta \geq \delta_\epsilon \), (17) is at most \( \epsilon \). The corollary then follows from (15) and the continuity of probability measures, since \( \mathbb{P}(R(S) \leq \delta) \to 1 \) as \( \delta \to \infty \).

Finally, we state a well-known result from the theory of Gaussian processes (see Adler [1] Inequality 2.4) which will be critical to our proof.

Lemma 3. Let \( \mathcal{X}(t) \) denote any centered, continuous Gaussian process, and \( T \) any bounded interval of \( \mathbb{R}^+ \). Let \( \sigma_T^2 = \sup_{t \in T} \mathbb{E}[\mathcal{X}^2(t)] \), and suppose \( \sigma_T^2 < \infty \). Then for all \( \epsilon > 0 \), there exists \( M_\epsilon \), depending only on \( \mathcal{X}(t), T, \) and \( \epsilon \), s.t. for all \( x > M_\epsilon \),
\[
\mathbb{P}\left( \sup_{t \in T} \mathcal{X}(t) > x \right) \leq \exp\left( -\left( 2\sigma_T^2 \right)^{-1} - \epsilon x^2 \right).
\]

We now complete the proof of Theorem 1.

Proof [Proof of Theorem 1] It follows from Corollary 4 and a union bound that for all $B > 0$, \( \limsup_{n \to \infty} \mathbb{P}\left( Q^n_B(\infty) \geq n \right) \) is at most
\[
\sum_{k=0}^{\infty} \mathbb{P}\left( \sup_{k \leq t \leq k+1} (\mathcal{Z}(t) - B \mu t) \geq \frac{B}{2} \right).
\]

Note that for any fixed $k \geq 0$, \( \mathbb{P}\left( \sup_{k \leq t \leq k+1} (\mathcal{Z}(t) - B \mu t) \geq \frac{B}{2} \right) \) equals
\[
\mathbb{P}\left( (\mathcal{Z}(k) - \frac{B}{4} \mu k) + \sup_{k \leq t \leq k+1} ((\mathcal{Z}(t) - B \mu t) - (\mathcal{Z}(k) - B \mu k) - \frac{B}{4} \mu k) \geq \frac{B}{2} \right).
\]
Let \( \mu' \overset{\Delta}{=} \min(\mu, 1) \). It then follows from the stationary increments property of \( Z(t) \) and a union bound that (18) is at most

\[
\sum_{k=0}^{\infty} \mathbb{P}\left( Z(k) \geq \frac{B}{4} \mu'(k+1) \right) + \sum_{k=0}^{\infty} \mathbb{P}\left( \sup_{0 \leq t \leq 1} Z(t) \geq \frac{B}{4} \mu'(k+1) \right).
\]

(19)

(20)

Since \( Z(0) = 0 \), and \( \frac{B}{4} \mu'(k+1) \geq \frac{B}{4} \mu'k \), (19) is at most

\[
\sum_{k=1}^{\infty} \mathbb{P}\left( k^{-\frac{1}{2}} Z(k) \geq \frac{B}{4} \mu'k^{\frac{1}{2}} \right).
\]

It follows from Theorem 6.(iv) that \( c \overset{\Delta}{=} \sup_{k \geq 1} \mathbb{E}\left[ \left( k^{-\frac{1}{2}} Z(k) \right)^2 \right] < \infty \) is a finite, strictly positive constant depending only on \( A \) and \( S \). Let \( \Phi \) denote the cumulative distribution function of a standard normally distributed r.v., \( \phi \) the corresponding probability density function, and let us define \( \Phi^c = 1 - \Phi \). Note that for all \( x \geq 1 \),

\[
\Phi^c(x) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} \exp\left( -\frac{y^2}{2} \right) dy \leq (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} y \exp\left( -\frac{y^2}{2} \right) dy = (2\pi)^{-\frac{1}{2}} \exp\left( -\frac{x^2}{2} \right).
\]

Combining the above, we find that (19) is at most

\[
\sum_{k=1}^{\infty} \exp\left( -\frac{\mu'^2}{32c} B^2 k \right) = \frac{\exp\left( -\frac{\mu'^2}{32c} B^2 \right)}{1 - \exp\left( -\frac{\mu'^2}{32c} B^2 \right)}.
\]

(21)

We now bound (20). Let \( c_2 \overset{\Delta}{=} \sup_{t \in [0, 1]} \mathbb{E}[Z^2(t)] \). It is easily verified that \( c_2 < \infty \). It follows from Lemma 3 that there exists \( B_0 < \infty \), depending only on \( A \) and \( S \), s.t. \( B \geq B_0 \) implies that for all \( k \geq 1 \),

\[
\mathbb{P}\left( \sup_{0 \leq t \leq 1} Z(t) > \frac{B}{4} \mu'k \right) \leq \exp\left( - (4c_2)^{-1}(\frac{B}{4} \mu'k)^2 \right).
\]

It follows that (20) is at most

\[
\sum_{k=1}^{\infty} \exp\left( -\frac{\mu'^2}{64c_2} B^2 k^2 \right) \leq \sum_{k=1}^{\infty} \exp\left( -\frac{\mu'^2}{64c_2} B^2 k \right) \leq \frac{\exp\left( -\frac{\mu'^2}{64c_2} B^2 \right)}{1 - \exp\left( -\frac{\mu'^2}{64c_2} B^2 \right)}.
\]

(22)

Using (21) to bound (19) and (22) to bound (20) completes the proof. \( \square \)
6. Proof of Bound for Probability of Delay as $B \to 0$

In this section we complete the proof of Theorem 2. It will be useful to first introduce some notions from asymptotic analysis, to aid in discussing how various quantities scale.

**Definition 1 (Asymptotic notation).** We say that a function $f(x)$ is

- $O_0(x)$ if $\limsup_{x \to 0} x^{-1} f(x) < \infty$, $O(x)$ if $\limsup_{x \to \infty} x^{-1} f(x) < \infty$;
- $\Omega_0(x)$ if $\liminf_{x \to 0} x^{-1} f(x) > 0$, $\Omega(x)$ if $\liminf_{x \to \infty} x^{-1} f(x) > 0$;
- $\Theta_0(x)$ if $f(x)$ is $O_0(x)$ and $\Omega_0(x)$, $\Theta(x)$ if $f(x)$ is $O(x)$ and $\Omega(x)$.

Now, note that the analysis used to treat the case $B \to \infty$, i.e. letting $\delta = \infty$, will not suffice here, as that approach would only be able to show that the s.s.p.d. is bounded from above by $1 - \exp(\epsilon B^2) = 1 - \Theta_0(B^2)$ for, while the “correct” scaling is actually $1 - \Theta_0(B)$. The underlying reason for this is that when $\delta = \infty$, $\eta$ must be chosen in such a way that stability is maintained, i.e. $\eta \in [0, B)$. In that case, the bound of Corollary 4 requires a stochastic process (i.e. $Z(t)$) with drift $-\Theta_0(B)$ to be within $\Theta_0(B)$ of zero. As $B \to 0$, this results in the aforementioned $1 - \Theta_0(B^2)$ behavior. Instead, we will select a non-trivial value for $\delta$. This will allow us to select a value for $\eta$ which is independent of $B$, as when $\delta < \infty$, Theorem 5 yields non-trivial bounds even when the choice of $\eta$ drives the queue into instability (i.e. $\eta > B$). In this case, our bounding system (in the H-W regime) then behaves like a stochastic process with drift $-\Theta_0(B)$ on $[0, t - \delta]$, and drift on the order of $\eta$ on $[t - \delta, t]$. As the probability that this process takes a “moderate value” (e.g. 3, independent of $B$) at time $t - \delta$ scales like $\Theta_0(B)$, while the probability that the process then drifts from this moderate value to below 0 on $[t - \delta, t]$ is just some constant depending on $\eta$ and $\delta$ (independent of $B$), we derive the desired bound of $1 - \epsilon B$. To formalize these arguments, we will rely on a careful analysis of the behavior of $Z(t) - Bt$ and its supremum and hitting times. We begin by reviewing several results from the theory of Gaussian processes for later use in our analysis.

6.1. Slepian’s lemma

We first review the well-known Slepian’s lemma (see Gordon [27]) for comparing the suprema of multivariate Gaussian r.v.s., as we will need these results for analyzing the supremum of $Z(t)$ by comparing to simpler processes. We formally state a particular variant of Slepian’s lemma, proven by Gordon [27] Theorem 1.1.

**Lemma 4.** For $k \geq 1$, let $(X_1, \ldots, X_k), (Y_1, \ldots, Y_k)$ denote two zero-mean multivariate Gaussian r.v.s, each in $\mathbb{R}^k$. Further suppose that $E[X_i^2] = E[Y_i^2], i = 1, \ldots, k,$ and $E[X_iX_j] \geq E[Y_iY_j]$ for all $i, j \in \{1, \ldots, k\}$. Then for all vectors $(z_1, \ldots, z_k) \in \mathbb{R}^k$,

$$
\mathbb{P}\left( \bigcap_{i=1,\ldots,k} \{X_i \leq z_i\}\right) \geq \mathbb{P}\left( \bigcap_{i=1,\ldots,k} \{Y_i \leq z_i\}\right).
$$

We now restate Lemma 4 for continuous Gaussian processes over an interval, which follows from Lemma 4, continuity, and the separability of $\mathbb{R}$ (see Piterbarg [45]).

**Corollary 5.** Let $T \overset{\Delta}{=} [t_1, t_2]$ denote any closed interval of $\mathbb{R}^+$. Let $\mathcal{X}(t)$ and $\mathcal{Y}(t)$ denote any two continuous zero-mean Gaussian processes s.t. $E[\mathcal{X}^2(t)] = E[\mathcal{Y}^2(t)]$ for all $t \in T$, and $E[\mathcal{X}(s)\mathcal{X}(t)] \geq E[\mathcal{Y}(s)\mathcal{Y}(t)]$ for all $s, t \in T$. Let $g(t)$ denote any function which is continuous on $T$. Then

$$
\mathbb{P}\left( \sup_{t \in T} (\mathcal{X}(t) - g(t)) \leq 0 \right) \geq \mathbb{P}\left( \sup_{t \in T} (\mathcal{Y}(t) - g(t)) \leq 0 \right).
$$

We also state another variant of Lemma 4 for Gaussian processes, which will be useful in our analysis.
Corollary 6. Let $T^1 \overset{\Delta}{=} [t_1, t_2]$, $T^2 \overset{\Delta}{=} [t_2, t_3]$ denote any closed intervals of $\mathbb{R}^+$ which intersect at exactly one point. Let $\mathcal{X}(t)$ denote any continuous zero-mean Gaussian processes s.t. $\mathbb{E} [\mathcal{X}(s)\mathcal{X}(t)] \geq 0$ for all $s, t \geq 0$. Let $g(t)$ denote any function which is continuous on $T^1 \cup T^2$. Then

$$\mathbb{P} \left( \sup_{t \in T^1 \cup T^2} (\mathcal{X}(t) - g(t)) \leq 0 \right) \geq \mathbb{P} \left( \sup_{t \in T^1} (\mathcal{X}(t) - g(t)) \leq 0 \right) \times \mathbb{P} \left( \sup_{t \in T^2} (\mathcal{X}(t) - g(t)) \leq 0 \right).$$

Proof The proof is deferred to the appendix. $\square$

6.2. Properties of Brownian motion, the Ornstein-Uhlenbeck process, and the three-dimensional Bessel process.

In this subsection we review several properties of Brownian motion (B.m.), the Ornstein-Uhlenbeck (O.U.) process, and the three-dimensional (3-D) Bessel process. In our analysis, we will construct a process whose supremum bounds that of $Z(t) - Bt$ (in an appropriate sense) by taking a weighted combination of independent B.m. and O.U. processes, and showing that the associated covariance structures satisfy the conditions of Slepian’s lemma. For this reason, we will need some relevant results pertaining to the supremum and hitting times of these processes, which we now state.

6.2.1. Brownian motion For $b > 0$, let $\mathcal{B}^b(t)$ denote a B.m initialized to $b$; namely, the continuous Gaussian process s.t. $\mathbb{E} [\mathcal{B}^b(t)] = b, V [\mathcal{B}^b(s), \mathcal{B}^b(t)] = s$ for all $0 \leq s \leq t$. We now state several basic properties of B.m, and refer the reader to Borodin and Salminen [4] for details. For a stochastic process $Z(t)$, let $\tau_Z^a$ denote the first hitting time of $Z(t)$ to $a$, where $\tau_Z^a = \infty$ if no such time exists.

Lemma 5. B.m has the following properties.

(i) For all $t \geq 0$, $\mathbb{P} \left( \sup_{0 \leq s \leq t} \mathcal{B}^b(s) > x \right) = 2\Phi (xt^{-\frac{1}{2}})$.

(ii) For all $c_1, c_2 > 0$, $\mathbb{P} (\tau_{\mathcal{B}^b}^{c_1} < \tau_{\mathcal{B}^b}^{c_2}) = \frac{c_2}{c_1 + c_2}$. 

(iii) For all $c, x > 0$, $\mathbb{P} \left( \sup_{t \geq 0} (\mathcal{B}^b(t) - ct) > x \right) = \exp (-2cx)$.

6.2.2. Ornstein-Uhlenbeck process For any $\rho > 0$, let $\mathcal{U}^\rho(t)$ denote the centered, stationary O.U. process whose correlations decay exponentially (over time) at rate $\rho$. Namely, $\mathcal{U}^\rho(t)$ is the continuous Gaussian process s.t. $\mathbb{E} [\mathcal{U}^\rho(t)] = 0, V [\mathcal{U}^\rho(s), \mathcal{U}^\rho(t)] = \exp \left( - \rho (t-s) \right)$ for all $0 \leq s \leq t$. For a review of the basic properties of O.U processes (e.g. existence, continuity), we refer the reader to Doob [17]. The main result w.r.t. O.U. processes that we will need in our analysis is a law of the iterated logarithm (i.i.m). We now state a particular variant, which follows from a more general result for continuous, stationary, Gaussian processes proven in Marcus [44].

Lemma 6. For any fixed $\rho > 0$, one may construct $\mathcal{U}^\rho(t)$ on the same probability space as an a.s. finite r.v. $Z$, whose distribution depends only on $\rho$, s.t. $|\mathcal{U}^\rho(t)| \leq Z + 2\log^\frac{3}{2} (t+1)$ for all $t \geq 0$ w.p.1.

6.2.3. Three-dimensional Bessel process For any $b > 0$, let $\mathcal{S}^b(t)$ denote the 3-D Bessel process initialized to $b$. We now formally define $\mathcal{S}^b(t)$ as the solution to a certain stochastic integral equation. The stochastic integral equation

$$X_t = b^2 + 3t + 2 \int_0^t |X_s|^\frac{1}{2} dB_s$$

has a unique strong solution $\mathcal{S}^b(t)$, which is non-negative; we refer the reader to the survey paper of G"{o}ing-Jaeschke and Yor [26] for details. Then $\mathcal{S}^b(t)$, the 3-D Bessel process initialized to $b$, is defined as $(\mathcal{S}^b(t))^\frac{1}{2}$. The 3-D Bessel processes will be useful in our analysis, since it has the same distribution as a B.m. conditioned to hit one level before another, an object which will arise when bounding the probability of certain events. In particular, it is proven by Williams [62], and restated by Pitman [47] Proposition 1.1, that
Lemma 7. For any fixed $0 < b < c < \infty$, the conditional distribution of the r.v.

$$\tau_{gb}^c$$
given \( \{ \tau_{gb}^c < \tau_{gb}^0 \} \)

is identical to the distribution of the r.v. \( \tau_{gb}^c \). Also, the conditional distribution of the process

$$B^b(t)_{0 \leq t \leq \tau_{gb}^c}$$
given \( \{ \tau_{gb}^c < \tau_{gb}^0 \} \)

is identical to the distribution of the process \( S^b(t)_{0 \leq t \leq \tau_{gb}^c} \).

We will also need an additional technical result w.r.t. the hitting times of \( S^b(t) \). Although these hitting times are well-studied (see Kent [37], Getoor and Sharpe [25], Pitman and Yor [46], Byczkowski and Ryznar [8], Byczkowski et al. [9]), we include a proof for completeness.

Lemma 8. For all \( b, \delta > 0 \), there exists \( T, \epsilon \in (0, \infty) \), depending only on \( b \) and \( \delta \), s.t. \( M > T \) implies \( \mathbb{P}(\tau_{gb}^b < \epsilon M^2) \leq \delta \).

Proof The proof is deferred to the appendix. \( \square \)

We will also need several other technical results w.r.t. \( S^b(t) \), for use in our analysis. It is well-known that \( S^0(t) \), namely the 3-D Bessel process initialized to 0, is distributed as \( (\sum_{i=1}^3 B^0_i(t)^2)^{1/2} \), namely the radial distance process of a 3-D B.M. More generally, an elegant construction for \( S^0(t) \) is given by Williams [62], where it is shown that for \( b > 0 \), \( S^b(t) \) is distributed as the ‘gluing together’ of two B.M.s initialized to \( b \), and a 3-D Bessel process initialized to 0. We now make this more precise. Let \( \{ U_b, \mu \geq 0 \} \) denote a set of independent uniformly distributed r.v.s, where \( U_b \) has the uniform distribution on \([0, b]\). Let \( \{ S^b_i(t), b \geq 0 \} \) denote a set of independent 3-D Bessel processes, where \( S^b_i(t) \) is initialized to \( b \). Suppose \( \{ B^b_i(t), b \in \mathbb{R}, i \geq 1 \}, \{ U_b, \mu \geq 0 \} \), and \( \{ S^b_i(t), b \geq 0 \} \) are mutually independent. Then it is proven by Williams [62] Theorem 3.1 that

Lemma 9. For \( b > 0 \), define

\[
\mathcal{X}(t) \triangleq \begin{cases} 
B^b_1(t) & 0 \leq t < \tau_{gb}^U, \\
B^b_2(\tau_{gb}^U + \tau_{gb}^U - t) & \tau_{gb}^U \leq t < \tau_{gb}^U + \tau_{gb}^U, \\
S^b_2(t - \tau_{gb}^U - \tau_{gb}^U) + b & \tau_{gb}^U + \tau_{gb}^U \leq t < \infty.
\end{cases}
\]

Then the distribution of the process \( \mathcal{X}(t) \) is identical to the distribution of the process \( S^b(t) \).

We will also need a l.i.l for 3-D Bessel processes, and we now state a particular variant, which follows from Hambly et al. [29] Theorem 2 and continuity.

Theorem 7. For any \( b, \epsilon > 0 \), one may construct \( S^b(t) \) on the same probability space as an a.s. finite r.v. \( Z \), whose distribution depends only on \( b \) and \( \epsilon \), s.t. \( S^b_i(t) \geq t^{2-\epsilon} - Z \) for all \( t \geq 0 \).

6.3. Renewal theory

We now review some results from renewal theory, which will be necessary in analyzing the covariance structure of \( Z(t) \). Let \( N^c(t) \) denote an equilibrium renewal process with renewal distribution \( S \), and \( N^c(t) \) denote an ordinary renewal process with renewal distribution \( S \). Let \( C_1 \triangleq \mu c^2, C_2 \triangleq \frac{4}{3} \mu^2 \mathbb{E}[S^2] + \frac{4}{3} \mu^4 (\mathbb{E}[S^2])^2 \), and \( C_3 \triangleq \mu^6 \mathbb{E}[S^2] \). Also, let \( f(t) \triangleq \mathbb{V}[N^c(t)] - C_1 t \). Then

Lemma 10. \( f(0) = 0 \), and \( sup_{t \geq 0} |f(t)| \leq C_2 \), \( f(t) \) is Lipschitz, namely \( |f(t+h) - f(t)| \leq C_3 h \) for all \( t, h \geq 0 \). Also, \( \mathbb{V}[N^c(s), N^c(t)] = C_1 s + \frac{4}{3} (f(s) + f(t) - f(t-s)) \) for all \( s, t \geq 0 \).
Proof That \( f(0) = 0 \) is trivial. That \( \sup_{t \geq 0} |f(t)| \leq C_2 \) follows from Daley [16] Equation 1.15. We now prove that \( |f(t + h) - f(t)| \leq C_3 h \) for all \( t, h \geq 0 \). Noting that \( \mathbb{E}[N^o(t)] \) is monotone and bounded on compact sets and thus integrable, it follows from Daley [16] Equation 1.4 that

\[
f(t) = 2\mu \int_0^t \left( \mathbb{E}[N^o(s)] + 1 - \mu s \right) - \frac{1}{2} (1 + c_3^2) \right) ds. \tag{24}
\]

It is proven in Lorden [39] that for all \( s \geq 0 \), one has \( 0 \leq \mathbb{E}[N^o(s)] + 1 - \mu s \leq \mu^2 \mathbb{E}[S^2] \), and it follows that

\[
\left| \left( \mathbb{E}[N^o(s)] + 1 - \mu s \right) - \frac{1}{2} (1 + c_3^2) \right| \leq \frac{1}{2} \mu^2 \mathbb{E}[S^2].
\]

Combining with (24) completes the proof.

We now prove the final assertion of the lemma.

\[
V[N^c(s), N^c(t)] = -\frac{1}{2} \left( \mathbb{E}[N^c(t) - N^c(s)]^2 - \mathbb{E}[N^c(s)]^2 - \mathbb{E}[N^c(t)]^2 \right) - \mu^2 st
\]

\[
= \frac{1}{2} \left( V[N^c(t)] + V[N^c(s)] - V[N^c(t - s)] \right),
\]

with the final equality following from stationary increments. The assertion then follows from definitions, completing the proof of the lemma. \( \square \)

We conclude this subsection by showing that the covariance of the number of renewals at different times of an equilibrium renewal process is always non-negative.

Lemma 11. \( \mathbb{E}[N^c(s)N^c(t)] - \mu^2 st \geq 0 \) for all \( s, t \geq 0 \). Furthermore, \( V[Z(s), Z(t)] \geq 0 \) for all \( s, t \geq 0 \).

Proof The proof is deferred to the appendix. \( \square \)

6.4. Bounding the covariance of \( Z(t) \)

In this subsection, we compare the covariance of \( Z(t) \) to that of a combination of a B.m. and an O.U. process. Let \( e \triangleq \exp(1) \), and \( \epsilon_0 \triangleq \left( 2(e - 2) \right)^{-1} \). It can be easily verified that \( \exp(-\epsilon_0) < \frac{1}{2} \). Let \( C_4 \triangleq \mu c_3^2 + C_1 \), and \( M \triangleq \frac{8c_2 + C_4 + C_1}{C_4 \left( \frac{1}{2} \exp(-\epsilon_0) \right)} \). Let \( U^M_1(t) \) denote a realization of the process \( U^M(t) \), independent of \( \{ \mathcal{B}^0_i(t), b \in \mathbb{R}, i \geq 1 \} \). Let us define a new Gaussian process \( \mathcal{W}(s) \) on \( [M, \infty) \). Note that \( MC_4 + f(s) > 0 \) for all \( s \geq 0 \) by the construction of \( M \), since \( |f(s)| \leq C_2 \) for all \( s \geq 0 \) by Lemma 10. We define

\[
\mathcal{W}(s) \triangleq C_4^\frac{1}{2} \left( 1 - \frac{M}{s} \right)^\frac{1}{2} \mathcal{B}^0_1(s) + \left( MC_4 + f(s) \right)^\frac{1}{2} U^M_1(s), s \geq M + 1.
\]

That \( \mathcal{W}(t)_{t \geq M+1} \) is a Gaussian process follows from the fact that sums of Gaussian processes are Gaussian processes, and a Gaussian process multiplied by a deterministic function of time is a Gaussian process. That \( \mathcal{W}(t)_{t \geq M+1} \) is continuous follows from the continuity of \( \mathcal{B}^0_1(t) \) and \( U^M_1(t) \), combined with the fact that \( f(s) \) is Lipschitz by Lemma 10. We now prove that

Theorem 8. For all \( s \geq M + 1 \) and \( t \geq s \), \( V[\mathcal{W}(s)] = V[Z(s)] \), and \( V[Z(s), Z(t)] \geq V[\mathcal{W}(s), \mathcal{W}(t)] \).

Proof The proof is deferred to the appendix. \( \square \)
6.5. Proof of Theorem 2
In this subsection, we complete the proof of Theorem 2.

Proof [Proof of Theorem 2] Suppose w.l.o.g. that $B \leq 1$. For a r.v. $Z$, let $m(Z)$ denote any median of $Z$. Since $E[S^2] < \infty$ implies $E(R(S)) < \infty$, let us define $\delta_0 \triangleq 2M + 1 + m(R(S))$. Since $D(t)$ is w.p.1 a continuous Gaussian process, $\sup_{0 \leq t \leq \delta_0} D(t)$ is a.s. finite, and we may define $\eta_0 \triangleq 4\left(m(\sup_{0 \leq t \leq \delta_0} D(t)) + 1\right)$. It follows from Theorem 5 and taking complements that

$$\liminf_{n \to \infty} \mathbb{P}\left(\left(Q^n_B(\infty) - n\right)m^{-\frac{1}{2}} \leq 0\right) \geq \mathbb{P}\left(\left(E_{B, \infty}^{\delta_0, \eta_0}(0)\right)^c\right).$$

It then follows from Lemma 11, Corollary 6, and the fact that $\mathbb{P}(R(S) \leq \delta_0) \geq \frac{1}{2}$, that $\mathbb{P}\left(\left(E_{B, \infty}^{\delta_0, \eta_0}(x)\right)^c\right)$ is at least

$$\mathbb{P}\left(\sup_{0 \leq t \leq \delta_0} (Z(t) + (\eta_0 - B)\mu t) \leq \frac{1}{2}\eta_0\right) \times \mathbb{P}\left(\sup_{t \geq \delta_0} (Z(t) - B\mu t) \leq -\eta_0\mu\delta_0\right).$$

We first bound (26). It follows from a straightforward union bound, and the independence of $A(t)$ and $D(t)$, that (26) is at least

$$\mathbb{P}\left(\sup_{0 \leq t \leq \delta_0} (A(t) + \eta_0\mu t) \leq \frac{1}{4}\eta_0\right) \times \mathbb{P}\left(\sup_{0 \leq t \leq \delta_0} D(t) \leq \frac{1}{4}\eta_0\right).$$

We now bound (27). We note that this is where the primary difficulty lies. We would like to show that (27) is $\Omega(1)$. However, there is no closed form available for the distribution of $\sup_{t \geq \delta_0}(Z(t) - B\mu t)$. Thus we will instead bound this supremum by comparing to $\sup_{t \geq \delta_0}(W(t) - B\mu t)$, i.e. the combination of a B.m. and O.U. process, whose supremum we can get a handle on using known results for the B.m. and O.U. processes. In particular, it follows from Theorem 8, Corollary 5, and the fact that $\delta_0 \geq M + 1$ that (27) is at least

$$\mathbb{P}\left(\sup_{t \geq \delta_0} \left(C_4^\delta \left(1 - \frac{M}{t}\right)^\frac{1}{2} B^\delta_4(t) + (MC_4^f(t))^\frac{1}{2} U^\delta_1(t) - B\mu t\right) \leq -\eta_0\mu\delta_0\right).$$

Note that $\sup_{t \geq \delta_0} (MC_4^f(t))^\frac{1}{2} \leq (MC_4^f + C_2)^\frac{1}{2}$ by Lemma 10. Thus it follows from Lemma 6 and the independence of $B^\delta_4(t)$ and $U^\delta_1(t)$ that we may construct an a.s. finite r.v. $Z_1$, whose distribution depends only on $A$ and $S$, on the same probability space as $B^\delta_4(t)$, s.t. $Z_1$ and $B^\delta_4(t)$ are independent, and (29) is at least

$$\mathbb{P}\left(\sup_{t \geq \delta_0} \left(C_4^\delta \left(1 - \frac{M}{t}\right)^\frac{1}{2} B^\delta_4(t) + Z_1 + 2(MC_4 + C_2)^\frac{1}{2} \log^\frac{1}{2}(t + 1) - B\mu t\right) \leq -\eta_0\mu\delta_0\right).$$

Since $Z_1$ is a.s. finite, $\delta_0 \geq 2M$ implies $(1 - \frac{M}{t})^\frac{1}{2} \geq \frac{1}{2}$, and $\log^\frac{1}{2}(t + 1) \leq t^\frac{1}{2} + 8$ for all $t \geq 0$, it follows that there exists a finite constant $C_5 \in (1, \infty)$, depending only on the distributions of $A$ and $S$, s.t. (30) is at least

$$\frac{1}{2} \mathbb{P}\left(\sup_{t \geq \delta_0} B^\delta_4(t) + C_5 t^\frac{1}{2} - C_5^{-1} Bt \leq -C_5\right).$$
where we choose to work with $t^{\frac{1}{4}}$ instead of $\log^2(t + 1)$ as a technical convenience, i.e. the same
difficulties arise in either case. Indeed, although Theorem 6.(iv) dictates that the variance of $Z(t)$
scales asymptotically linearly in $t$, to bound the supremum we must explicitly account for the
lower-order behavior, leading to the “correction” term $C_5t^{\frac{1}{4}}$. By Theorem 7, i.e. the l.i.l for the
3-D Bessel process, there exists an a.s. finite non-negative r.v. $Z_2$, whose distribution depends only
on the distributions of $A$ and $S$, s.t. $S_1(t) \geq t^{\frac{1}{4}} - Z_2$ for all $t \geq 0$. Let $C_6 \triangleq (m(Z_2) + C_0^2)$, and
$H_1 \triangleq C_5(1 + \delta_0^2) - C_6 + 1$. Then it follows from stationary increments, and the fact that $(x + y)^{\frac{1}{4}} \leq x^{\frac{1}{4}} + y^{\frac{1}{4}}$ for all $x, y \geq 0$, that (31) is at least
\[
\frac{1}{2}\mathbb{P}(B_1^0(\delta) \leq -H_1) \times \mathbb{P}\left(\sup_{t \geq 0} \left( B_1^0(t) + C_5 t^{\frac{1}{4}} - C_5^{-1} Bt \right) \leq -C_6 + 1 \right).
\] (32)
The symmetries of B.m. imply that (32) equals
\[
\mathbb{P}\left( \inf_{t \geq 0} \left( B_1^0(t) - C_5 t^{\frac{1}{4}} + C_5^{-1} Bt \right) \geq C_6 \right). \tag{33}
\]
We will now show that the correction term $-C_5 t^{\frac{1}{4}}$ does not change the fundamental qualitative
behavior of (33) as $B \to 0$, i.e. it still scales as $\Omega_0(B)$. We note that without the correction term,
(33) could be analyzed directly using known results about the supremum of B.m. with a linear
drift.

We bound (33) from below by conditioning on an appropriate event related to hitting times, and
then applying Lemma 7, i.e. the relationship between B.m. with appropriately conditioned hitting
times and the 3-D Bessel process. It follows from Lemma 8, applied with $b = 1$ and $\delta = \frac{1}{8}$, that
there exist finite constants $\epsilon, T > 0$, independent of $B$, s.t. for all $x \geq T$, one has
\[
\mathbb{P}(\tau_{s_1^0}^{xB^{-1}} < \epsilon x^2 B^{-2}) \leq \frac{1}{8}. \tag{34}
\]
Let $H_3 \triangleq TC_5 + 2C_5^4 \epsilon^{-\frac{1}{2}}$, and $H_4 \triangleq (2C_5^2)^{\frac{1}{4}}$, and note that $H_3 \geq T$, and $H_4 H_3^{-2} < \epsilon$. It thus follows
from (34) that
\[
\mathbb{P}\left( \tau_{s_1^0}^{H_3 B^{-1}} < H_4 B^{-2} \right) \leq \frac{1}{8}. \tag{35}
\]
Let
\[
\mathcal{E}_1 \triangleq \left\{ \inf_{0 \leq t \leq \tau_{s_1^0}^{H_3 B^{-1}}} \left( B_1^0(t) - C_5 t^{\frac{1}{4}} + C_5^{-1} Bt \right) \geq C_6 \right\},
\]
and
\[
\mathcal{E}_2 \triangleq \left\{ \inf_{t \geq \tau_{s_1^0}^{H_3 B^{-1}}} \left( B_1^0(t) - C_5 t^{\frac{1}{4}} + C_5^{-1} Bt \right) \geq C_6 \right\}.
\]
Trivially, (33) equals $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)$. Let $\mathcal{E}_3 \triangleq \left\{ \tau_{s_1^0}^{H_4 B^{-1}} < \tau_{B_1^0}^{1} \right\}$. $\mathcal{E}_3$ will be useful in our analysis, as $\mathbb{P}(\mathcal{E}_3)$
is $\Theta_0(B)$ by Lemma 5.(i), and $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2)$ can be analyzed using the properties of the 3-D Bessel
process after conditioning on $\mathcal{E}_3$ and applying Lemma 7. Let us now define $\mathcal{E}_4 \triangleq \left\{ \tau_{s_1^0}^{H_3 B^{-1}} \geq H_4 B^{-2} \right\}$. $\mathcal{E}_4$ will be useful, since conditional on $\mathcal{E}_4$, $t \geq \tau_{s_1^0}^{H_3 B^{-1}}$ implies $C_5 t^{\frac{1}{4}} \leq \frac{1}{2} C_5^{-1} B t$, i.e. the term $C_5 t^{\frac{1}{4}}$ can
be safely ignored, and we can analyze $\mathbb{P}(\mathcal{E}_4|\mathcal{E}_3)$ using known properties of the 3-D Bessel process. In particular, let $\mathcal{E}_2^t \triangleq \left\{ \inf_{t \geq \tau_{s_1^2}} H_{s_1^2}^{-1} \left( B_1^0(t) + \frac{1}{2} C_5^{-1} B t \right) \geq C_6 \right\}$, it follows from the above and a union bound that

\[
\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq \mathbb{P}(\mathcal{E}_3) \times (\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2|\mathcal{E}_3)) \\
\geq \mathbb{P}(\mathcal{E}_3) \times (\mathbb{P}(\mathcal{E}_2|\mathcal{E}_3) - \mathbb{P}(\mathcal{E}_2 - \mathcal{E}_3)).
\]

We now bound $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2|\mathcal{E}_3)$. Let $\mathcal{E}_2'' \triangleq \left\{ \inf_{t \geq 0} \left( H_{s_1^2}^{-1} + B_1^0(t) + \frac{1}{2} C_5^{-1} B t \right) \geq C_6 \right\}$. It follows from the strong Markov property of B.m., and the fact that $C_5^{-1} B(t + \tau_{s_1^2}^{-1}) \geq C_5^{-1} B t$, that

\[
\mathbb{P}(\mathcal{E}_2|\mathcal{E}_3) \geq \mathbb{P}(\mathcal{E}_2|\mathcal{E}_3) \times \mathbb{P}(\mathcal{E}_2'').
\]

It follows from Lemma 5.(iii) and the symmetries of B.m. that

\[
\mathbb{P}(\mathcal{E}''_2) = 1 - \exp (-H_{s_1^2}^{-1} + BC_6 C_5^{-1}) \geq \frac{1}{2},
\]

where the final inequality follows from the fact that $-H_{s_1^2}^{-1} + BC_6 C_5^{-1} \leq -1$ by construction, and

\[
1 - e^{-1} \geq \frac{1}{2}.
\]

We now bound $\mathbb{P}(\mathcal{E}_1|\mathcal{E}_3)$. It follows from Lemma 7, which relates an appropriately conditioned B.m. to the 3-D Bessel process, that

\[
\mathbb{P}(\mathcal{E}_1|\mathcal{E}_3) = \mathbb{P}\left( \inf_{0 \leq t \leq \tau_{s_1^2}} (S_1^1(t) - C_5 t^{\frac{1}{4}} + C_5^{-1} B t) \geq C_6 \right) \\
\geq \mathbb{P}\left( \inf_{t \geq 0} (S_1^1(t) - C_5 t^{\frac{1}{4}}) \geq C_6 \right).
\]

Recall from Theorem 7, i.e. the l.i.l for the 3-D Bessel process, that there exists an a.s. finite non-negative r.v. $Z_2$ s.t. $S_1^1(t) \geq t^{\frac{3}{8}} - Z_2$ for all $t \geq 0$. It follows that (39) is at least

\[
\frac{1}{2} I \left( \left\{ \inf_{t \geq 0} (t^{\frac{3}{8}} - C_5 t^{\frac{1}{4}}) \geq m(Z_2) + C_6 \right\} \right).
\]

We now show that the indicator appearing in (40) evaluates to 1, from which it will follow that $\mathbb{P}(\mathcal{E}_1|\mathcal{E}_3) \geq \frac{1}{2}$. By construction, $m(Z_2) + C_6 = -C_5^3$. As $t^{\frac{3}{8}} \geq C_5 t^{\frac{1}{4}}$ for all $t \geq C_5^3$, it suffices to demonstrate that $t^{\frac{3}{8}} - C_5 t^{\frac{1}{4}} + C_5^3 \geq 0$ for all $t \in [0, C_5^3]$. The desired claim then follows from the fact that $C_5 t^{\frac{1}{4}} \leq C_5^3$ for all $t \in [0, C_5^3]$.

Combining the above with (37), (38) and (40), we conclude that

\[
\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2|\mathcal{E}_3) \geq \frac{1}{4}.
\]

We now bound $\mathbb{P}(\mathcal{E}_1' \cap \mathcal{E}_3)$, which by Lemma 7 equals

\[
\mathbb{P}(\tau_{s_1^2}^{-1} \geq H_4 B^{-2}) \leq \frac{1}{8}.
\]
where the final inequality follows from (35). Combining (29) - (33) with (36), we conclude that (33) is at least \( \frac{4}{8} \mathbb{P}(E_3) \). Combining with (26) - (27), (28), and (29) - (32), it follows that \( \mathbb{P}\left(\left( E_{B,\infty}^{\delta,n}(x) \right)^c \right) \) is at least

\[
\mathbb{P}\left( \sup_{0 \leq t \leq \delta_0} (A(t) + \eta_0 \mu t) \leq \frac{1}{4} \eta_0 \right) \times \mathbb{P}\left( \sup_{0 \leq t \leq \delta_0} D(t) \leq \frac{1}{4} \eta_0 \right)
\]

\[
\times \frac{1}{2} \mathbb{P}\left( E_1^{\delta_0}(\delta_0) \leq -H_1 \right)
\]

\[
\times \frac{1}{8} \mathbb{P}(E_3).
\]

It follows from Lemma 5.(ii) that \( \mathbb{P}(E_3) = H_3^{-1}B \). Noting that (43) - (44) are strictly positive numbers which depend only on \( A \) and \( S \), and combining with (25), completes the proof of the theorem. \( \square \)

7. Proof of Large Deviations Result

In this section we complete the proof of Theorem 3. Intuitively, we will relate the probability of there being fewer than \( n - x n^{1/2} \) jobs in the system at a large time \( t \) to the corresponding probability in the system where \( \Theta(|x|)n^{1/2} \) servers break down at time \( t - \delta \), for some moderate value of \( \delta \), and the remaining \( n - \Theta(|x|)n^{1/2} \) servers are kept busy by adding artificial arrivals on the remaining time horizon \([t - \delta, t]\). When translated to the H-W regime, this allows us to bound the desired probability from below by the probability that a Gaussian process, with drift \( \Theta(|x|) \), only grows moderately (i.e., not an amount proportional to \(|x|\)) over a time interval of length \( \delta \). Applying known results about Gaussian processes will then yield the desired result.

**Proof** [Proof of Theorem 3] Let us fix some \( B > 0 \), and \( x < -1 \). It follows from Theorem 6.(iv) that \( c_0 \overset{\Delta}{=} 16 \left( m + m \left( \sup_{t \geq 0} (D(t) - \frac{B}{2} \mu t) \right) + m \left( R(S) \right) + \mu^{-1} \right) \in (16, \infty) \). Let

\[
F_1 \overset{\Delta}{=} \left\{ \sup_{0 \leq t \leq c_0} (Z(t) + (c_0|x| - B) \mu t) \leq x + c_0|x| \mathbb{P}(R(S) \leq c_0) \right\},
\]

and

\[
F_2 \overset{\Delta}{=} \left\{ \sup_{t \geq c_0} (Z(t) - B \mu t + c_0^2|x| \mu) \leq x + c_0|x| \mathbb{P}(R(S) \leq c_0) \right\}.
\]

Then applying Theorem 5 with \( \delta = c_0, \eta = c_0|x| \) and taking complements, we find that

\[
\liminf_{n \to \infty} \mathbb{P}\left( (Q^n_B(\infty) - n)^{1/2} \leq x \right) \geq \mathbb{P}(F_1 \cap F_2).
\]

Let

\[
F_1^\prime \overset{\Delta}{=} \left\{ \sup_{0 \leq t \leq c_0} (A(t) + c_0|x| \mu t) \leq \frac{1}{8} c_0|x| \right\},
\]

and

\[
F_2^\prime \overset{\Delta}{=} \left\{ \sup_{t \geq c_0} (A(t) - \frac{1}{2} B \mu t + c_0^2 \mu |x|) \leq \frac{1}{8} c_0|x| \right\}.
\]

Then it follows from a straightforward union bound, and the fact that \( \mathbb{P}(R(S) \leq c_0) \geq \frac{1}{2}, \frac{1}{2} c_0 - 1 \geq \frac{1}{4} c_0 \), and \( \mathbb{P}\left( \sup_{t \geq 0} (D(t) - \frac{1}{2} B \mu t) \geq \frac{1}{8} c_0 \right) \leq \frac{1}{2}, \) that \( \mathbb{P}(F_1 \cap F_2) \geq \frac{1}{2} \mathbb{P}(F_1^\prime \cap F_2^\prime) \). Although an exact analysis of \( \mathbb{P}(F_1^\prime \cap F_2^\prime) \) in principle follows from known results about B.m. (see e.g. Shepp [52],
Borodin and Salminen [4], Boukai [6]), a simpler analysis suffices for our purposes, which we include for completeness. Let
\[ F_1' = \left\{ \sup_{0 \leq t \leq (16\mu)^{-1}} A(t) \leq 1 \right\} \bigcap \left\{ A((16\mu)^{-1}) \leq -c_0^2|x| - 1 \right\}, \]
and \[ F_2' = \left\{ \sup_{t \leq c_0} (A(t) - A(c_0) - \frac{1}{2}B\mu t) \leq 1 \right\}. \] Since \( \frac{1}{8}c_0|x| \geq 1 + \frac{1}{16}c_0|x| \geq 1 \), it follows from a straightforward union bound that \( F_1' \cap F_2' \subseteq F_1 \cap F_2' \). It follows from a slight modification of Corollary 6, the details of which we omit, and the strong Markov property of B.m. that
\[ \mathbb{P}(F_1' \cap F_2') \geq \mathbb{P} \left( \sup_{0 \leq t \leq (16\mu)^{-1}} A(t) \leq 1 \right) \times \mathbb{P} \left( A((16\mu)^{-1}) \leq -c_0^2|x| - 1 \right) \times \mathbb{P} \left( \sup_{t \geq 0} (A(t) - \frac{1}{2}B\mu t) \leq 1 \right). \]
Combining the above with Lemma 5.(iii) and the basic properties of B.m. completes the proof. □

8. Comparison to Other Bounds From the Literature
In this section we compare our results to known results for the GI/H\(_2\)/n, GI/D/n, and M/GI/∞ queues, showing that our main results are tight, in an appropriate sense.

8.1. GI/H\(_2\)/n queue
Let us define \( \alpha(x) \triangleq (1 + x\Phi(x)\phi^{-1}(x))^{-1} \). In Whitt [61], the authors prove the following, which generalizes (and corrects) the results for the GI/M/n queue given in Halfin and Whitt [28].

**Lemma 12.** Whitt [61] Suppose that \( \{Q^n_B, n \geq 1\} \) is a sequence of GI/H\(_2\)/n queues satisfying the H-W assumptions, i.e. with probability \( p \in (0, 1] \), S is exponentially distributed with mean \( (p\mu_S)^{-1} \), and with probability \( 1 - p \), S equals 0. Also suppose that \( \mathbb{E}[A^3] < \infty \). Let \( z \triangleq \frac{1}{2}p(c_A^2 + c_S^2) \). Then for all \( x \geq 0 \),
\[ \lim_{n \to \infty} \mathbb{P} \left( |Q^n_B(\infty) - n|^{-\frac{1}{2}} \geq x \right) = \alpha(Bz^{-\frac{1}{2}}) \exp(-Bpz^{-1}x), \]
and
\[ \lim_{n \to \infty} \mathbb{P} \left( |Q^n_B(\infty) - n|^{-\frac{1}{2}} \leq -x \right) = (1 - \alpha(Bz^{-\frac{1}{2}})) \frac{\Phi((B - x)z^{-\frac{1}{2}})}{\Phi(Bz^{-\frac{1}{2}})}. \]
It follows from Lemma 12 and a straightforward asymptotic analysis (the details of which we omit) that

**Corollary 7.** Under the same assumptions as Lemma 12,
(i) \( \lim_{B \to \infty} B^{-2} \log \left( \lim_{n \to \infty} \mathbb{P}(Q^n(\infty) \geq n) \right) = -\left( p(c_A^2 + c_S^2) \right)^{-1} > -\infty, \)
(ii) \( \lim_{B \to 0} B^{-1} \lim_{n \to \infty} \mathbb{P}(Q^n(\infty) < n) = \pi \left( p(c_A^2 + c_S^2) \right)^{-1} < \infty, \)
(iii) \( \lim_{x \to \infty} x^{-2} \log \left( \lim_{n \to \infty} \mathbb{P} \left( |Q^n(\infty) - n|^{-\frac{1}{2}} \leq -x \right) \right) = -\left( p(c_A^2 + c_S^2) \right)^{-1} < 0. \)
Thus in this case, we conclude that all of our main results capture the correct qualitative scaling of the relevant quantities.
8.2. GI/D/n queue
In Jelenkovic et al. [34], the authors prove the following.

**Lemma 13.** Jelenkovic et al. [34] Suppose that \( \{Q^n_B, n \geq 1\} \) is a sequence of GI/D/n queues satisfying the H-W assumptions, i.e. the processing times are distributed as some strictly positive constant. Also suppose that \( \mathbb{E}[A^2] < \infty \). Let \( \{X_i, i \geq 1\} \) denote a set of i.i.d. normally distributed r.v.s with mean \(-B\) and variance \( c_A^2 > 0 \). Let \( S_i = \sum_{k=1}^i X_k, i \geq 0 \). Then for all \( x \in \mathbb{R} \),

\[
\lim_{n \to \infty} \mathbb{P}\left( (Q^n_B(\infty) - n) n^{-\frac{1}{2}} \geq x \right) = \mathbb{P}(\sup_{i \geq 1} S_i \geq x).
\]

We will now prove that Lemma 13 implies the following.

**Corollary 8.** Under the same assumptions as Lemma 13,

\( (i) \) \( \lim_{B \to -\infty} B^{-2} \log \left( \lim_{n \to \infty} \mathbb{P}(Q^n(\infty) \geq n) \right) = -(2c_A)^{-1} > -\infty, \)

\( (ii) \) \( \lim_{B \to -\infty} B^{-1} \lim_{n \to \infty} \mathbb{P}(Q^n(\infty) < n) = 2^\frac{1}{2} c_A^{-1} < \infty, \)

\( (iii) \) \( \lim_{x \to -\infty} x^{-2} \log \left( \lim_{n \to \infty} \mathbb{P}\left( (Q^n(\infty) - n) n^{-\frac{1}{2}} \leq -x \right) \right) = -(2c_A)^{-1} < 0. \)

**Proof** (ii) is proven in Jelenkovic et al. [34] Section 4.2. We now prove (i). Let \( B' \overset{\Delta}{=} Bc_A^{-1} \). Then it follows from Lemma 13, and the asymptotic analysis of the supremum of the Gaussian random walk given in Janssen and Van Leeuwaarden [33], in particular Equation 2.18, that

\[
\lim_{B \to -\infty} B^{-2} \log \left( \lim_{n \to \infty} \mathbb{P}(Q^n(\infty) \geq n) \right) = \lim_{B \to -\infty} B^{-2} \log \left( 1 - \exp\left( -(2\pi B^2)^{-\frac{1}{2}} \exp(-\frac{1}{2}B^2) \right) \right).
\]

Combining with a straightforward asymptotic analysis completes the proof.

Finally, we complete the proof of (iii). Note that for all \( x \in \mathbb{R} \),

\[
\mathbb{P}(S_i \geq x) \leq \mathbb{P}(\sup_{i \geq 1} S_i \geq x) \leq \mathbb{P}\left( \sup_{i \geq 1} (c_A B^0(t) - Bt) \geq x \right).
\]

The desired claim then follows from the strong Markov property of B.m., and the basic asymptotics of \( \Phi \). \( \square \)

Again, we conclude that all of our main results capture the correct qualitative scaling of the relevant quantities.

8.3. M/GI/∞ lower bound
Suppose that \( Q^n_B \) is an M/GI/n queue satisfying the H-W assumptions. Let \( Z_{n,B} \) denote a Poisson r.v. with mean \( \lambda_{n,B} \). Then it follows from a straightforward infinite-server lower bound, and the well-known properties of the steady-state infinite server queue (see Takacs [55]), that for all \( x \in \mathbb{R}^+ \),

\[
\mathbb{P}(Q^n_B(\infty) \leq n - xn^\frac{1}{2}) \leq \mathbb{P}(Z_{n,B} \leq n - xn^\frac{1}{2}).
\]

It follows from the Central Limit Theorem that for all \( x \in \mathbb{R}^+ \), \( \lim_{n \to \infty} \mathbb{P}(Z_{n,B} \leq n - xn^\frac{1}{2}) = \Phi(B - x) \), and we conclude the following.

**Lemma 14.** Suppose that \( \{Q^n_B, n \geq 1\} \) is a sequence of M/GI/n queues satisfying the H-W assumptions. Then

\[
\lim_{B \to -\infty} \liminf_{n \to \infty} B^{-2} \log \left( \liminf_{n \to \infty} \mathbb{P}(Q^n_B(\infty) \geq n) \right) \geq -\frac{1}{2} > -\infty,
\]

and

\[
\lim_{x \to -\infty} x^{-2} \log \left( \limsup_{n \to \infty} \mathbb{P}\left( (Q^n(\infty) - n) n^{-\frac{1}{2}} \leq -x \right) \right) \leq -\frac{1}{2} < 0.
\]

It follows that in this setting, Theorem 1 and Theorem 3 are again tight. Interestingly, the infinite server lower-bound does not seem to yield any information about the tightness of Theorem 2, since \( \Phi(B) \geq \frac{1}{2} \) for all \( B \geq 0 \).
9. Conclusion and Open Questions

In this paper, we studied the FCFS GI/GI/n queue in the H-W regime, deriving bounds on the s.s.p.d. and number of idle servers. In particular, we proved that there exists $\epsilon > 0$, depending on the inter-arrival and processing time distributions, s.t. the s.s.p.d. is bounded from above by $\exp(-\epsilon B^2)$ as the associated excess parameter $B \to \infty$; and by $1 - \epsilon B$ as $B \to 0$. We also proved that the tail of the steady-state number of idle servers has a Gaussian decay. Furthermore, we used known results for the special cases of $H^*_2$ and deterministic processing times, as well as the $M/GI/\infty$ queue, to prove that our bounds are tight, in an appropriate sense. Our main proof technique was the derivation of new stochastic comparison bounds for the FCFS GI/GI/n queue, which are of a structural nature, and build on the recent work of Gamarnik and Goldberg [21].

This work leaves many interesting directions for future research. The explicit limits for the s.s.p.d. as $B \to 0$ and $B \to \infty$ for $H^*_2$ and deterministic processing times are suggestive of some fascinating patterns, and it is an intriguing open problem to understand the nature of these limits for more general distributions. A reasonable place to start may be with distributions for which we at least know that the relevant sequence of normalized steady-state queue-length distributions converges weakly to some limit, e.g. processing times with finite support. It is also interesting to note that for $H^*_2$ and deterministic processing times, $\lim_{B \to \infty} B^{-2} \log(p_B)$ is actually equal to $\lim_{x \to \infty} x^{-2} \log(F_{idle,B}(x))$ - to what extent such a relationship may hold in general is unknown. Furthermore, these limits may fit into the broader context of insensitivity results in queueing, in which various limits depend on the relevant distributions only through limited information (e.g. first two moments), and it is open to investigate such connections.

We believe that the stochastic comparison techniques developed in this paper, and the original work of Gamarnik and Goldberg [21], have the potential to shed insight on many other queueing models. Indeed, one may view our methods as a step towards developing a calculus of stochastic-comparison type bounds for parallel server queues, in which one derives bounds by composing structural modifications (e.g. adding jobs, removing servers, adding servers, etc.) over time. Three particularly interesting systems to which one might try to apply these methods are queues with abandonments, queues with heavy-tailed processing times, and networks of queues, which arise in various applied settings. We note that the setting of abandonments and networks are particularly interesting from a stochastic comparison standpoint, as these systems may exhibit certain non-monotonicities (e.g. adding a job may cause other jobs to leave the system sooner), and developing a better understanding of when such systems can be compared will likely require the creation of new tools. Developing stochastic comparison techniques to derive a matching lower bound for Theorem 2 also remains an open challenge. It is also interesting to study the extent to which limiting results such as Theorems 1 - 3 apply for any fixed $n$, and refer the reader to Gamarnik and Stolyar [23] for some results along these lines.

On a final note, the different qualitative behavior we observe w.r.t. the s.s.p.d. as $B \to 0$ and $B \to \infty$ fits into the broader theme of the H-W regime as a “transition” between a system which behaves like an infinite-server queue ($B \to \infty$) and a single-server queue ($B \to 0$). In the H-W regime, this transition was previously formalized for the case of Markovian processing times by Gamarnik and Goldberg [20] and Knessl and Leeuwaarden [38], by proving that there exists $B^* \approx 1.85722$ s.t. the spectral measure of the underlying Markov chain had no jumps (like the single-server queue) for all $B \in (0, B^*)$, and had at least one jump (like the infinite-server queue) for all $B > B^*$. Furthermore, in the work of Knessl and Leeuwaarden [38], a certain notion of “convergence” of the spectral measure to that of the infinite-server queue, as $B \to \infty$, is demonstrated. It is an interesting open question to more generally formalize these “limits within limits” of the H-W regime.
10. Appendix
10.1. Proof of Corollary 6

Proof [Proof of Corollary 6] Consider some set $t_1, \ldots, t_k$ of times belonging to the interval $T^1$, and a set $t_{k+1}, \ldots, t_{2k}$ of times belonging to the interval $T^2$. Let $(Y_1, \ldots, Y_{2k}) \in \mathbb{R}^{2k}$ denote the zero-mean multivariate Gaussian r.v. s.t. $E[Y_i^2] = E[X^2(t_i)], i = 1, \ldots, 2k$, $E[Y_i Y_j] = E[X(t_i)X(t_j)], 1 \leq i \leq j \leq k$, $E[Y_i Y_j] = E[X(t_i)X(t_j)], k+1 \leq i \leq j \leq 2k$, and $E[Y_i Y_j] = 0$ otherwise (i.e. $i \in \{1, \ldots, k\}$, $j \in \{k+1, \ldots, 2k\}$). That such a multivariate Gaussian exists follows from the fact that $(Y_1, \ldots, Y_k)$ is distributed as $(X(t_1), \ldots, X(t_k))$, $(Y_{k+1}, \ldots, Y_{2k})$ is distributed as $(X(t_{k+1}), \ldots, X(t_{2k}))$, and $(Y_1, \ldots, Y_k)$ is independent of $(Y_{k+1}, \ldots, Y_{2k})$. Then we may apply Lemma 4 to find that

$$
\mathbb{P}\left( \bigcap_{i=1}^{2k} \{X(t_i) \leq g(t_i)\} \right) \geq \mathbb{P}\left( \bigcap_{i=1}^{k} \{Y_i \leq g(t_i)\} \right)
= \mathbb{P}\left( \bigcap_{i=1}^{k} \{Y_i \leq g(t_i)\} \right) \times \mathbb{P}\left( \bigcap_{i=k+1}^{2k} \{Y_i \leq g(t_i)\} \right).
$$

The corollary then follows by continuity and separability, by letting $\{t_i, i = 1, \ldots, 2k\}$ become dense in $T^1 \cup T^2$. \hfill \Box

10.2. Proof of Lemma 8

In this subsection we complete the proof of Lemma 8.

Proof [Proof of Lemma 8] Note that the event $\{\tau_{b_0}^M < \epsilon M^2\}$ equals the event $\{\sup_{0 \leq t \leq \epsilon M^2} \mathcal{S}_1^b(t) > M\}$. It follows from Lemma 9 that $U_b, B_1^b(t), B_2^b(t), \mathcal{S}_1^b(t)$, and $\mathcal{S}_1^0(t)$ can be constructed on the same probability space s.t.

$$
sup_{0 \leq t \leq \epsilon M^2} \mathcal{S}_1^b(t) \leq sup_{0 \leq t \leq \epsilon M^2} U_b(t) + sup_{0 \leq t \leq \epsilon M^2} B_2^b(t) + sup_{0 \leq t \leq \epsilon M^2} \mathcal{S}_1^0(t) + b.
$$

Since $\tau_{b_0}^U \leq \tau_{b_0}^0$, and $B_i^b(t)$ is distributed as $b + B_i^0(t)$ for all $i$, it follows from the above, a union bound, and the fact that $M > 6b$ that $\mathbb{P}\left( \sup_{0 \leq t \leq \epsilon M^2} \mathcal{S}_1^b(t) > M\right)$ is at most

$$
2\mathbb{P}\left( \tau_{b_0}^M > \epsilon M^2\right) + 2\mathbb{P}\left( \sup_{0 \leq t \leq \epsilon M^2} B_1^0(t) > \frac{M}{6} \right) + \mathbb{P}\left( \sup_{0 \leq t \leq \epsilon M^2} \mathcal{S}_1^0(t) > \frac{M}{6} \right).
$$

(46)

We now bound the first term of (46). It follows from the symmetries of B.m. that $\tau_{b_0}^0$ has the same distribution as $\tau_{b_0}^0$. It follows from Lemma 5.(i) and a union bound that the first term of (46) is at most

$$
2\mathbb{P}\left( \tau_{b_0}^b > \epsilon M^2\right) = 2\mathbb{P}\left( \sup_{0 \leq t \leq \epsilon M^2} B_1^0(t) < b\right)
= 2 \left( 1 - 2\Phi\left(b(\epsilon M^2)^{-\frac{1}{2}}\right) \right) \leq 4 \int_0^{b(\epsilon M^2)^{-\frac{1}{2}}} \phi(x) dx \leq 4(2\pi)^{-\frac{1}{2}} b e^{-\frac{1}{2} M^{-1}},
$$

(47)

where the final inequality follows from the fact that $\phi(x) \leq (2\pi)^{-\frac{1}{2}}$ for all $x$.

We now bound the second term of (46), which again by Lemma 5.(i) is at most

$$
4\Phi\left(\frac{M}{6}(\epsilon M^2)^{-\frac{1}{2}}\right) \leq 4 \exp \left( \frac{1}{2} - \frac{1}{6} \epsilon^{-\frac{1}{2}} \right),
$$

(48)
We now bound the third and final term of (46). Recall that \( S_i^1(t) \) has the same distribution as \( \left( \sum_{i=1}^{3} (B_i^0(t))^2 \right)^{\frac{1}{2}} \), and note that \( \left( \sum_{i=1}^{3} (B_i^0(t))^2 \right)^{\frac{1}{2}} \leq \sum_{i=1}^{3} |B_i^0(t)|. \) It thus follows from a union bound, Lemma 5.(i), and a Chernoff bound that that the third term of (46) is at most

\[
6\mathbb{P}\left( \sup_{0 \leq t \leq M^2} B_i^0(t) > \frac{M}{18} \right) \leq 12 \exp\left( \frac{1}{2} - \frac{1}{18} \epsilon^{-\frac{1}{2}} \right). \tag{49}
\]

Using (47),(48), and (49) to bound (46) completes the proof. \( \Box \)

10.3. Proof of Theorem 8

In this subsection we complete the proof of Theorem 8. Let us fix some \( s \geq M + 1 \) and \( t \geq s \). Recall that \( N^s(t)(N^o(t)) \) denotes an equilibrium (ordinary) renewal process with renewal distribution \( S \), \( C_1 = \mu e_2^3, C_2 = \frac{1}{2} \mu^3 \mathbb{E}[S^3] \), \( C_3 = \mu^3 \mathbb{E}[S^2] \), \( C_4 = \mu e_2^3 + C_1, e = \exp(1), e_0 = (2(e - 2))^{-1}, M = \frac{8(C_2 + C_3 + C_4)}{C_4(\frac{1}{2} - \exp(1))} \), and \( f(t) = V[N^s(t) - C_1 t] \). We begin by establishing several technical preliminaries.

**Lemma 15.**

(i) \( (1 - \frac{M}{s})^{\frac{1}{2}} (1 - \frac{M}{t})^{\frac{1}{2}} s \leq s - M + \frac{(t-s)M}{2t}. \)

(ii) \( (1 - \frac{M}{s})^{\frac{1}{2}} (1 - \frac{M}{t})^{\frac{1}{2}} s \leq s - \frac{M}{2}. \)

(iii) \( (MC_4 + f(s))^{\frac{1}{2}} (MC_4 + f(t))^{\frac{1}{2}} \leq MC_4 + f(s) + \frac{C_4}{2}(t-s). \)

(iv) \( (MC_4 + f(s))^{\frac{1}{2}} (MC_4 + f(t))^{\frac{1}{2}} \leq MC_4 + C_2. \)

(v) \( V[Z(s), Z(t)] \geq C_4 s + f(s) - C_3(t-s). \)

(vi) \( V[Z(s), Z(t)] \geq C_4 s - 3C_2. \)

**Proof (i):**

\[
(1 - \frac{M}{s})^{\frac{1}{2}} (1 - \frac{M}{t})^{\frac{1}{2}} s = (s - M) \left( \frac{1 - \frac{M}{s}}{1 - \frac{M}{t}} \right)^{\frac{1}{2}} = (s - M) \left( 1 + \frac{(t-s)M}{st - tM} \right)^{\frac{1}{2}} \leq (s - M) \left( 1 + \frac{(t-s)M}{2(s-M)t} \right).
\]

where the final inequality follows from the fact that

\[
(1 + x)^{\frac{1}{2}} \leq 1 + \frac{1}{2} x \quad \text{for all} \quad x \geq -1. \tag{50}
\]

(ii):

\[
(1 - \frac{M}{s})^{\frac{1}{2}} (1 - \frac{M}{t})^{\frac{1}{2}} s \leq (1 - \frac{M}{s})^{\frac{1}{2}} s \leq s - \frac{M}{2} \quad \text{by (50)}.
\]

(iii):

\[
(MC_4 + f(s))^{\frac{1}{2}} (MC_4 + f(t))^{\frac{1}{2}} = (MC_4 + f(s)) (1 + \frac{f(t) - f(s)}{MC_4 + f(s)})^{\frac{1}{2}}
\]
\[
\leq (MC_4 + f(s)) \left(1 + \frac{|f(t) - f(s)|}{2(MC_4 + f(s))}\right) \quad \text{by (50)}
\]
\[
\leq (MC_4 + f(s)) \left(1 + \frac{C_3(t - s)}{2MC_4 + f(s)}\right)
\]
\[
= MC_4 + f(s) + \frac{C_3}{2}(t - s),
\]
where the final inequality follows from Lemma 10.

(iv): Follows from Lemma 10.

(v): It follows from Lemma 10 and the triangle inequality that
\[
V[Z(s), Z(t)] = V[A(s), A(t)] + V[D(s), D(t)]
\]
\[
= C_4s + \frac{1}{2}(f(s) + f(t) - f(t - s))
\]
\[
\geq C_4s + f(s) - \frac{|f(t) - f(s)| + |f(t - s)|}{2}
\]
\[
\geq C_4s + f(s) - C_3(t - s).
\]

(vi): It follows from Lemma 10 that
\[
V[Z(s), Z(t)] = C_4s + \frac{1}{2}(f(s) + f(t) - f(t - s))
\]
\[
\geq C_4s - 3C_2.
\]

\]

We now complete the proof of Theorem 8.

**Proof** [Proof of Theorem 8] First, note that
\[
V[W(s)] = V[C_4^2 (1 - \frac{M}{s})^\frac{3}{2}B^m_1(s)] + V[(MC_4 + f(s))^\frac{3}{2}U^M(s)]
\]
\[
= C_4(1 - \frac{M}{s})s + MC_4 + f(s) = V[Z(s)].
\]

In general,
\[
V[W(s), W(t)] = V[C_4^2 (1 - \frac{M}{s})^\frac{3}{2}B^m_1(s), C_4^2 (1 - \frac{M}{t})^\frac{3}{2}B^m_1(t)]
\]
\[
+ V[(MC_4 + f(s))^\frac{3}{2}U^M_1(s), (MC_4 + f(t))^\frac{3}{2}U^M_1(t)]
\]
\[
= C_4(1 - \frac{M}{s})^\frac{3}{2}(1 - \frac{M}{t})^\frac{3}{2}s
\]
\[
+ (MC_4 + f(s))^\frac{3}{2}(MC_4 + f(t))^\frac{3}{2} \exp \left(-M(t - s)\right). \quad \text{(51)}
\]

We now treat two cases. First, suppose \( t - s \leq \epsilon_0 M^{-1} \). Then (51) is at most \( C_4 \left( s - M + \frac{(t - s)M}{2t} \right) \) by Lemma 15.(i). It follows from Lemma 15.(iii) that (52) is at most
\[
(MC_4 + f(s) + \frac{C_3}{2}(t - s)) \exp \left(-M(t - s)\right).
\]
It follows from a simple Taylor series expansion that for all \( y \in [0, \epsilon_0] \), one has \( \exp(-y) \leq 1 - \frac{y}{2} \). Combining the above, we conclude that (52) is at most

\[
(MC_4 + f(s) + \frac{C_3}{2} (t-s)) (1 - \frac{M}{2} (t-s))
\]

\[
= MC_4 + f(s) + \frac{C_3}{2} (t-s) - \frac{M^2 C_4}{2} (t-s) - \frac{Mf(s)}{2} (t-s) - \frac{MC_3}{4} (t-s)^2
\]

\[
\leq MC_4 + f(s) + \left( \frac{C_3}{2} + \frac{MC_2}{2} - \frac{M^2 C_4}{2} \right) (t-s),
\]

where the final inequality follows since \(|f(s)| \leq C_2\) by Lemma 10. Combining our bounds for (51) and (52), we conclude that

\[
V[\mathcal{W}(s), \mathcal{W}(t)] \leq C_4 s - MC_4 + \frac{MC_4}{2t} (t-s)
\]

\[
+ MC_4 + f(s) + \left( \frac{C_3}{2} + \frac{MC_2}{2} - \frac{M^2 C_4}{2} \right) (t-s)
\]

\[
\leq C_4 s + f(s) + \left( \frac{C_3}{2} + \frac{MC_2}{2} + \frac{MC_4}{2t} - \frac{M^2 C_4}{2} \right) (t-s)
\]

\[
\leq C_4 s + f(s) + \frac{2M(C_2 + C_3 + C_4) - C_4}{2} M^2 \right) (t-s),
\]

where the final inequality follows since \( M, t \geq 1 \). Combining with Lemma 15.(v), we conclude that \( V[\mathcal{Z}(s), \mathcal{Z}(t)] - V[\mathcal{W}(s), \mathcal{W}(t)] \) is at least

\[
\left( \frac{C_4}{2} M^2 - 2M(C_2 + C_3 + C_4) - C_3 \right) (t-s)
\]

\[
\geq M \left( \frac{C_4}{2} M - 4(C_2 + C_3 + C_4) \right) (t-s)
\]

\[
= 4M(C_2 + C_3 + C_4) \left( \frac{1}{2} \exp(-\epsilon_0) - 1 \right) (t-s) \geq 0,
\]

where the final equality follows from the definition of \( M \).

Alternatively, suppose \( t - s \geq \epsilon_0 M^{-1} \). Then (51) is at most \( C_4 (s - \frac{M}{2}) \) by Lemma 15.(ii), and (52) is at most \( (MC_4 + C_2) \exp(-\epsilon_0) \) by Lemma 15.(iv). Combining the above, we find that

\[
V[\mathcal{W}(s), \mathcal{W}(t)] \leq C_4 (s - \frac{M}{2}) + (MC_4 + C_2) \exp(-\epsilon_0)
\]

\[
= C_4 s + MC_4 \left( 1 + \frac{C_2}{MC_4} \exp(-\epsilon_0) - \frac{1}{2} \right).
\]

(53)

Thus by Lemma 15.(vi), \( V[\mathcal{Z}(s), \mathcal{Z}(t)] - V[\mathcal{W}(s), \mathcal{W}(t)] \) is at least

\[
MC_4 \left( \frac{1}{2} - \left( 1 + \frac{C_2}{MC_4} \right) \exp(-\epsilon_0) - \frac{3C_2}{MC_4} \right)
\]

\[
= MC_4 \left( \frac{1}{2} - \exp(-\epsilon_0) - \left( \exp(-\epsilon_0) + 3 \right) C_2 (MC_4)^{-1} \right)
\]

\[
\geq MC_4 \left( \frac{1}{2} - \exp(-\epsilon_0) - \frac{8(C_2 + C_3 + C_4)}{MC_4} \right) \geq 0,
\]

(54)

by the definition of \( M \). Combining the above completes the proof. □
10.4. Proof of Lemma 11
In this subsection we complete the proof of Lemma 11.

Proof [Proof of Lemma 11] Let \{X_k, k \geq 1\} denote the ordered sequence of renewal intervals in process \(N^c(t)\). Then from definitions,

\[
V[N^c(s), N^c(t)] = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s) I(\sum_{k=1}^{j} X_k \leq t)] - \mu^2 st. \tag{55}
\]

Note that for \(j \leq i\), one has that

\[
\mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s) I(\sum_{k=1}^{j} X_k \leq t)] = \mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s)]
\geq \mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s)] \mathbb{E}[I(\sum_{k=1}^{j} X_k \leq t)]. \tag{56}
\]

Alternatively, suppose \(j \geq i + 1\). Let \(Y_1 \triangleq \sum_{k=1}^{i} X_k, Y_2 \triangleq \sum_{k=i+1}^{j} X_k\), and \(Y_3 \triangleq t - Y_1\). Then \(Y_1\) and \(Y_2\) are independent, \(Y_2\) and \(Y_3\) are independent, and

\[
\mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s) I(\sum_{k=1}^{j} X_k \leq t)] = \mathbb{E}[I(Y_1 \leq s) I(Y_1 + Y_2 \leq t)]
= \mathbb{E}[I(Y_3 \geq t - s) I(Y_3 \geq Y_2)]
= \mathbb{E}[I(Y_2 \geq t - s) I(Y_3 \geq Y_2) + I(Y_2 < t - s) I(Y_3 \geq t - s)]. \tag{57}
\]

Let \(Y_a, Y_b\) denote two r.v.s, each distributed as \(Y_3\), where \(Y_a, Y_b\) are mutually independent. Then by linearity of expectation, (57) equals

\[
\mathbb{E}[I(Y_2 \geq t - s) I(Y_a \geq Y_2) + I(Y_2 < t - s) I(Y_b \geq Y_2)]
\geq \mathbb{E}[I(Y_2 \geq t - s) I(Y_a \geq Y_2) I(Y_b \geq t - s)]
+ \mathbb{E}[I(Y_2 < t - s) I(Y_a \geq Y_2) I(Y_b \geq Y_2)]
= \mathbb{E}[I(Y_a \geq Y_2) I(Y_b \geq t - s)]
= \mathbb{E}[I(Y_1 + Y_2 \leq t)] \mathbb{E}[I(Y_1 \leq s)]. \tag{58}
\]

Combining (55) - (58), we find that

\[
V[N^c(s), N^c(t)] \geq \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s)] \mathbb{E}[I(\sum_{k=1}^{j} X_k \leq t)] - \mu^2 st
\]

\[
= \mathbb{E}[\sum_{i=1}^{\infty} I(\sum_{k=1}^{i} X_k \leq s)] \mathbb{E}[\sum_{j=1}^{\infty} I(\sum_{k=1}^{j} X_k \leq t)] - \mu^2 st
= \mathbb{E}[N^c(s)] \mathbb{E}[N^c(t)] - \mu^2 st = 0,
\]

completing the proof of the first part of the lemma. The second part of the lemma then follows from definitions, since \(V[B(s), B(t)]\) is also non-negative. \(\square\)

Acknowledgements
The author would like to thank Dimitris Bertsimas, Ton Dieker, David Gamarnik, and Kavita Ramanan for their helpful discussions and insights.
References

[1] R.J. Adler. An introduction to continuity, extrema, and related topics for Gaussian processes. *Inst. Math. Statist. Lecture Notes - Monograph Series*, 12, 1990.

[2] S. Asmussen. *Applied Probability and Queues*, 2nd ed. Springer, 2003.

[3] P. Billingsley. *Convergence of Probability Measures*. 1999.

[4] A. Borodin and P. Salminen. *Handbook of Brownian Motion - Facts and Figures*, 2nd edition. Birkhauser Verlag, 2002.

[5] A.A. Borovkov. Some limit theorems in the theory of mass service, II. *Theor. Probability Appl.*, 10:375–400, 1965.

[6] B. Boukai. An explicit expression for the distribution of the supremum of Brownian motion with a change-point. *Purdue University Technical Report*, (88-40), 1988.

[7] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, S. Zeltyn, L. Zhao, and S. Haipeng. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100:36–50, 2005.

[8] T. Byczkowski and M. Ryznar. Hitting distribution of geometric Brownian motion. *Studia. Math.*, 173(1):19–38, 2006.

[9] T. Byczkowski, J. Malecki, and M. Ryznar. Hitting times of Bessel processes. *Preprint*, 2010.

[10] C.S. Chang, J.A. Thomas, and S.H. Kiang. On the stability of open networks: a unified approach by stochastic dominance. *Queueing Systems*, 15:239–260, 1994.

[11] D. Cox. *Renewal Theory*. Methuen and Co., 1970.

[12] JG Dai and S. He. Many-server queues with customer abandonment: numerical analysis of their diffusion models. *arXiv preprint arXiv:1103.0347*, 2011.

[13] JG Dai and S. He. Many-server queues with customer abandonment: A survey of diffusion and fluid approximations. *Journal of Systems Science and Systems Engineering*, pages 1–36, 2012.

[14] JG Dai, S. He, and T. Tezcan. Many-server diffusion limits for G/Ph/n+ GI queues. *The Annals of Applied Probability*, 20(5):1854–1890, 2010.

[15] JG Dai, AB Diicker, and Xuefeng Gao. Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. 2013.

[16] D. J. Daley. Bounds for the variance of certain stationary point processes. *Stochastic Processes and their Applications*, 7(3):255 – 264, 1978.

[17] J.L. Doob. The Brownian movement and stochastic equations. *The Annals of Mathematics*, 43(2):351–369, 1942.

[18] J.L. Doob. The elementary Gaussian processes. *Ann. Math. Statist.*, 15(3):229–282, 1944.

[19] A.K. Erlang. *On the rational determination of the number of circuits*. The Copenhagen Telephone Company, Copenhagen, 1948.

[20] D. Gamarnik and D.A. Goldberg. On the rate of convergence to stationarity of the M/M/n queue in the Halfin-Whitt regime. *The Annals of Applied Probability*. To appear. URL http://www.e-publications.org/ims/submission/index.php/AAP/user/submissionFile/6985?confirm=eb6c4c77

[21] D. Gamarnik and D.A. Goldberg. Steady-state GI/G1/n queue in the Halfin-Whitt regime. *The Annals of Applied Probability*. To appear. URL http://www.e-publications.org/ims/submission/index.php/AAP/user/submissionFile/9531?confirm=f8d5cb3c

[22] D. Gamarnik and P. Momeilovic. Steady-state analysis of a multi-server queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 40:548–577, 2008.

[23] D. Gamarnik and A.L. Stolyar. Multiclass multiserver queueing system in the halfin–whitt heavy traffic regime: asymptotics of the stationary distribution. *Queueing Systems*, pages 1–27, 2012.

[24] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.
[25] R.K. Getoor and M.J. Sharpe. Excursions of Brownian motion and Bessel processes. *Probability Theory and Related Fields*, 47:83–106, 1979.

[26] A. Going-jaeschke and M. Yor. A survey and some generalizations of Bessel processes. *Bernoulli*, 9: 313–349, 2003.

[27] Y. Gordon. Some moment inequalities for Gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

[28] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[29] B.M. Hambly, G. Kersting, and A.E. Kyprianou. Law of the iterated logarithm for oscillating random walks conditioned to stay positive. *Stoch. Proc. Appl.*, 108:327–343, 2003.

[30] J. Huang, H. Zhang, and J. Zhang. Diffusion approximations for queueing systems with customer abandonment. *arXiv preprint arXiv:1208.5552*, 2012.

[31] I. Ibragimov and Y. Rozanov. *Gaussian random processes*. 1978.

[32] D. Jagerman. Some properties of the Erlang loss function. *Bell System Techn. J.*, 53(3):525–551, 1974.

[33] AJEM Janssen and JSH Van Leeuwaarden. Cumulants of the maximum of the gaussian random walk. *Stochastic Processes and Their Applications*, 117(12):1928–1959, 2007.

[34] P. Jelenkovic, A. Mandelbaum, and P. Momcilovic. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems: Theory and Applications*, 47(1-2):53–69, 2004.

[35] W. Kang and K. Ramanan. Asymptotic approximations for stationary distributions of many-server queues with abandonment. *The Annals of Applied Probability*, 22(2):477–521, 2012.

[36] H. Kaspi and K. Ramanan. SPDE limits of many-server queues. *The Annals of Applied Probability*, 23(1):145–229, 2013.

[37] J. Kent. Some probabilistic properties of Bessel functions. *Annals of Probability*, 6(6):760–770, 1978.

[38] C. Knessl and JSH Van Leeuwaarden. Transient behavior of the Halfin-Whitt diffusion. *Stochastic Processes and their Applications*, 121(7):1524–1545, 2011.

[39] G. Lorden. On excess over the boundary. *The Annals of Mathematical Statistics*, 41(2):520–527, 1970.

[40] A. Mandelbaum and P. Momcilovic. Queues with many servers: the virtual waiting-time process in the QED regime. *Math. Oper. Res.*, 33(3):561–586, 2008.

[41] A. Mandelbaum and P. Momčilović. Queues with many servers and impatient customers. *Mathematics of Operations Research*, 37(1):41–65, 2012.

[42] A. Mandelbaum and S. Zeltyn. Call centers with impatient customers: many-server asymptotics of the M/M/n+ G queue. *Queueing Systems*, 51(3):361–402, 2005.

[43] M. Marcus and J. Rosen. *Markov Processes, Gaussian Processes, and Local Times*. 2006.

[44] M.B. Marcus. Upper bounds for the asymptotic maxima of continuous Gaussian processes. *The Annals of Mathematical Statistics*, 43(2):522–533, 1972.

[45] V. Piterbarg. *Asymptotic Methods in the theory of Gaussian processes and fields*. American Mathematical Society, 1996.

[46] J. Pitman and M. Yor. *Stochastic Integrals*. Springer, New York, 1981.

[47] J.W. Pitman. One-dimensional Brownian motion and the three-dimensional Bessel process. *Advances in Applied Probability*, 7(3):511–526, 1975.

[48] A. Puhalski and M. Reiman. The multiclass GI/PH/n queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 32(3):564–595, 2000.

[49] A.A. Puhalski and J. Reed. On many-server queues in heavy traffic. *Ann. Appl. Prob.*, 20(1):129–195, 2010.

[50] J. Reed. The G/GI/n queue in the Halfin-Whitt regime. *Annals of Applied Probability*, 19(6):2211–2269, 2009.
[51] J. Reed and T. Tezcan. Hazard rate scaling of the abandonment distribution for the GI/M/n+ GI queue in heavy traffic. *Operations Research*, 60(4):981–995, 2012.

[52] L.A. Shepp. The joint density of the maximum and its location for a Wiener process with drift. *Journal of Applied probability*, 16:423–427, 1979.

[53] A.V. Skorokhod. Stochastic equations for diffusions in a bounded region. *Theory Probab. Appl.*, 6:264–274, 1961.

[54] D. Stoyan. *Comparison methods for queues and other stochastic models*. Wiley, 1983.

[55] L. Takacs. *Introduction to the theory of queues*. Oxford University Press, New York, 1962.

[56] A. Ward. Asymptotic analysis of queueing systems with reneging: A survey of results for fifo, single class models. *Surveys in Operations Research and Management Science*, 17(1):1–14, 2012.

[57] A. Weerasinghe. Diffusion approximations for G/M/n+ GI queues with state-dependent service rates. *Mathematics of Operations Research*, 2013.

[58] W. Whitt. Queues with superposition arrival processes in heavy traffic. *Stoch. Proc. App.*, 21:81–91, 1985.

[59] W. Whitt. *Stochastic Process Limits*. Springer, 2002.

[60] W. Whitt. A diffusion approximation for the G/GI/n/m queue. *Oper. Res.*, 52(6):922–941, November 2004. ISSN 0030-364X. doi: 10.1287/opre.1040.0136. URL http://dx.doi.org/10.1287/opre.1040.0136.

[61] W. Whitt. Heavy-traffic limits for the G/H2*/n/m queue. *Math. Oper. Res.*, 30(1):1–27, February 2005. ISSN 0364-765X. doi: 10.1287/moor.1040.0119. URL http://dx.doi.org/10.1287/moor.1040.0119.

[62] D. Williams. Path decomposition and continuity of local time for one-dimensional diffusions, I. *Proc. London Math. Soc.*, 28:738–768, 1974.

[63] Y. Zheng, N. Shroff, and P. Sinha. Heavy traffic limits for GI/H/n queues: Theory and application. 2012.