Submitted to the Annals of Applied Probability

STEADY-STATE GI/GI/n QUEUE IN THE HALFIN-WHITT REGIME

By David Gamarnik and David A. Goldberg

MIT and Georgia Institute of Technology

We consider the FCFS $GI/GI/n$ queue in the so-called Halfin-Whitt heavy traffic regime. We prove that under minor technical conditions the associated sequence of steady-state queue length distributions, normalized by $n^{\frac{1}{2}}$, is tight. We derive an upper bound on the large deviation exponent of the limiting steady-state queue length matching that conjectured by Gamarnik and Momcilovic in [16]. We also prove a matching lower bound when the arrival process is Poisson.

Our main proof technique is the derivation of new and simple bounds for the FCFS $GI/GI/n$ queue. 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. We further illustrate the utility of this methodology by deriving the first non-trivial bounds for the weak limit process studied in [37].

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, which can be described as critical with respect to the probability that an arriving customer has to wait for service. Namely, in this regime the stationary probability of wait is bounded away from both 0 and 1, as the number of servers grows. Although studied originally by Erlang [15] and Jagerman [22], the regime was formally introduced by Halfin and Whitt [18], who studied the $GI/M/n$ system (for large n) when the traffic intensity scales like $1 - Bn^{-\frac{1}{2}}$ for some strictly positive B. Namely, the parameter B controls how close to overloaded the system is in heavy traffic. They proved that under minor technical assumptions on the inter-arrival distribution, this sequence of $GI/M/n$ queueing models has the following properties:

(i) the steady-state probability that an arriving job finds all servers busy (i.e. the probability of wait) has a non-trivial limit;

(ii) the sequence of queueing processes, normalized by $n^{\frac{1}{2}}$, converges weakly to a non-trivial positive recurrent diffusion;

(iii) the sequence of steady-state queue length distributions, normalized by $n^{\frac{1}{2}}$, is tight and converges distributionally to the mixture of a point mass at 0 and an exponential distribution.

Furthermore, this steady-state probability of wait can be parametrized as a function of B, with larger values of B corresponding to smaller probabilities of wait. Similar weak convergence results under the H-W scaling were subsequently obtained for more general multi-server systems [32], [23], [28], [16], [17], [37] with the most general (single-class) results appearing in [37] (and follow-up papers [36],[34]). 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, with the exception of [18] (which treats exponential processing times), [23] (which treats

AMS 2000 subject classifications: Primary 60K25

Keywords and phrases: Many-server queues, large deviations, weak convergence, Gaussian process

deterministic processing times), [16] (which treats processing times with finite support), and [17] (which treats phase-type processing times and allows for abandonments and multi-class structure), all of the aforementioned results are for the associated sequence of normalized *transient* queue length distributions only, leaving many open questions about the associated *steady-state* queue length distributions.

In particular, in [16] it is shown for the case of processing times with finite support that the sequence of steady-state queue length distributions (normalized by $n^{\frac{1}{2}}$) is tight, and has a limit whose tail decays exponentially fast. The authors further prove that this exponential rate of decay (i.e. large deviation exponent) is $-2B(c_A^2+c_S^2)^{-1}$, where B is the spare capacity parameter, and c_A^2, c_S^2 are the squared coefficients of variation of the inter-arrival and processing time distributions. In [16] it was conjectured that this result should hold for more general processing time distributions. However, prior to this work no further progress on this question has been achieved.

In this paper we resolve the conjectures made in [16] with regards to (w.r.t.) tightness of the steady-state queue length, and take a large step towards resolving the conjectures made w.r.t. the large deviation exponent. We prove that as long as the inter-arrival and processing time distributions satisfy minor technical conditions (e.g. finite $2+\epsilon$ moments), the associated sequence of steady-state queue length distributions, normalized by $n^{\frac{1}{2}}$, is tight. Under the same minor technical conditions we derive an upper bound on the large deviation exponent of the limiting steady-state queue length matching that conjectured by Gamarnik and Momcilovic in [16]. We also prove a matching lower bound when the arrival process is Poisson.

Our main proof technique is the derivation of new and simple bounds for the FCFS $GI/GI/n$ queue. 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. Our upper and lower bounds also exhibit a certain duality relationship, and exemplify a general methodology which may be useful for analyzing a variety of queueing systems. We further illustrate the utility of this methodology by deriving the first non-trivial bounds for the weak limit process studied in [37].

We note that 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. Interestingly, such ideas were used in the original paper of Halfin and Whitt [18] to show tightness of the steady-state queue length for the $GI/M/n$ queue under the H-W scaling, but do not seem to have been used in subsequent works on queues in the H-W regime.

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 establish our general-purpose lower bounds for the queue length in a properly initialized FCFS $M/GI/n$ queue. In Section 5 we use our bounds to prove the tightness of the steady-state queue length when the system is in the H-W regime. In Section 6 we combine our bounds with known results about weak limits and the suprema of Gaussian processes to prove our large deviation results. In Section 7 we use our bounds to study the weak limit derived in [37]. In Section 8 we summarize our main results and comment on directions for future research. We include a technical appendix in Section 9.

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.) random variables (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 σ_A^2 and σ_S^2 denote the variance of A

and S, respectively. Let c_A^2 and c_S^2 denote the squared coefficient of variation (s.c.v.) of A and S, respectively.

We fix some excess parameter $B > 0$, and let $\lambda_n \stackrel{\Delta}{=} n - Bn^{\frac{1}{2}}$. For n sufficiently large to ensure $\lambda_n > 0$ (which is assumed throughout), let $Q^n(t)$ 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 $A\lambda_n^{-1}$ 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 in [18], as the traffic intensity in the nth 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.

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 $\{Q^n(t), n \geq 1\}$ 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 μ. 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) $\limsup_{t \downarrow 0} t^{-1}\mathbb{P}(S \leq t) < \infty$.
(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 now briefly discuss the various assumptions, commenting on both the reason for their inclusion and their restrictiveness. Condition (i) is necessary for several bounds from the literature relating to suprema of random walks (see [44]). Although we use this condition to prove tightness of the queue length in the H-W regime, all our intermediate results about weak limits and Gassian processes would also hold under only a second moment assumption (as opposed to $2 + \epsilon$).

Condition (iii) is necessary for several results from the literature relating to the weak convergence of scaled renewal processes (see [47], [49]). The condition is (for example) satisfied by any discrete distribution with no mass at zero, any continuous distribution with finite density at zero, and (more generally) any distribution function (d.f.) which is absolutely continuous in a neighborhood of zero (see the discussion in [49]). All our results other than those pertaining to the weak convergence of scaled renewal processes and/or the large deviation exponent of the queue length in the H-W regime would also hold without this assumption.

Condition (iv) is needed to sensibly discuss the relevant stationary measures. We refer the interested reader to [4] for an excellent discussion of sufficient conditions on A and S which ensure that (iv) holds, for example if the d.f. of A is continuous, or more generally has a "spread-out component" (see [4] for details). We note that our non-asymptotic transient bounds hold even without this condition.

We now state our main results. We begin by establishing the tightness of the steady-state queue length for the FCFS $GI/GI/n$ queue in the H-W regime.

THEOREM 1. *If A and S satisfy the H-W and T_0 assumptions, then the sequence $\{\left(Q^n(\infty) - n\right)^+ n^{-\frac{1}{2}}, n \geq 1\}$ is tight.*

In words, the queue length $\big(Q^n(\infty)-n\big)^+$ scales like $O(n^{\frac{1}{2}})$. Although we conjecture that the sequence $\{\big(Q^n(\infty)-n\big)^+ n^{-\frac{1}{2}}, n\geq 1\}$ has a unique weak limit (and thus converges weakly), our approach, which proves the weak convergence of certain bounding processes for the $GI/GI/n$ queue (but not the $GI/GI/n$ queue itself) is unable to establish this, and we leave the question of uniqueness as an interesting open problem.

We now establish an upper bound for the large deviation exponent of the limiting steady-state queue length for the FCFS $GI/GI/n$ queue in the H-W regime, and a matching lower bound when the arrival process is Poisson.

THEOREM 2. *Under the same assumptions as Theorem 1,*

$$\limsup_{x\to\infty} x^{-1}\log\left(\limsup_{n\to\infty}\mathbb{P}\Big(\big(Q^n(\infty)-n\big)^+ n^{-\frac{1}{2}}>x\Big)\right)\leq -2B(c_A^2+c_S^2)^{-1}.$$

If in addition A is an exponentially distributed r.v., namely the system is $M/GI/n$, then

$$\begin{aligned}&\lim_{x\to\infty} x^{-1}\log\left(\liminf_{n\to\infty}\mathbb{P}\Big(\big(Q^n(\infty)-n\big)^+ n^{-\frac{1}{2}}>x\Big)\right)\\ =\ &\lim_{x\to\infty} x^{-1}\log\left(\limsup_{n\to\infty}\mathbb{P}\Big(\big(Q^n(\infty)-n\big)^+ n^{-\frac{1}{2}}>x\Big)\right) \quad = \quad -2B(c_A^2+c_S^2)^{-1}.\end{aligned}$$

In words, Theorem 2 states that the tail of the limiting steady-state queue length is bounded from above by $\exp\big(-2B(c_A^2+c_S^2)^{-1}x+o(x)\big)$; and when the arrival process is Poisson, the tail of the limiting steady-state queue length is bounded from below by $\exp\big(-2B(c_A^2+c_S^2)^{-1}x-o(x)\big)$, where $o(x)$ is some non-negative function s.t. $\lim_{x\to\infty} x^{-1}o(x)=0$. Theorem 2 translates into bounds for the large deviation behavior of any weak limit of the sequence $\{\big(\mathrm{Q}^n(\infty)-n\big)^+ n^{-\frac{1}{2}}, n\geq 1\}$, where at least one weak limit exists by Theorem 1.

Note that the functional form of the exponent $-2B(c_A^2+c_S^2)^{-1}$ shows that the probability of large deviations is a decreasing function of the excess parameter B, and an increasing function of the squared coefficients of variation c_A^2, c_S^2. This is consistent at an intuitive level, since as B grows, the system becomes less loaded, which should decrease the the probability of large deviations. Similarly, as c_A^2 and c_S^2 grow, the system becomes more variable, which should increase the probability of large deviations.

Although we conjecture that $-2B(c_A^2+c_S^2)^{-1}$ should also be the correct large deviations exponent when A is non-Markovian, our lower-bounding proof technique relies on certain properties of the steady-state $M/GI/\infty$ queue which do not hold for the steady-state $GI/GI/\infty$ queue, and thus we leave such an extension as an open problem.

3. Upper bound. In this section, we prove a general upper bound for the FCFS $GI/GI/n$ queue, when properly initialized. The bound is valid for all finite n, and works in both the transient and steady-state (when it exists) regimes. Although we will later customize this bound to the H-W regime to prove our main results, we note that the bound is in no way limited to that regime. For a non-negative r.v. X with finite mean $\mathbb{E}[X]>0$, let $R(X)$ denote a r.v. distributed as the residual life distribution of X. Namely, for all $z\geq 0$,

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

Recall that associated with a 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, including the first, 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 equilibrium renewal process with renewal distribution A, with $A(t), \{N_i(t)\}$ mutually independent.

Let $\mathcal{Q}$ denote the FCFS $GI/GI/n$ queue with inter-arrival times drawn i.i.d. distributed as A, processing times drawn i.i.d. distributed as S, and the following initial conditions. For $i = 1, \ldots, n$, there is a single job initially being processed on server i, and the set of initial processing times of these n initial jobs is drawn i.i.d. distributed as $R(S)$. There are zero jobs waiting in queue, and the first inter-arrival time is distributed as $R(A)$, independent of the initial processing times of those jobs initially in system. We now establish an upper bound for $Q(t)$, the number in system at time t in $\mathcal{Q}$.

THEOREM 3. *For all $x > 0$, and $t \geq 0$,*

$$\mathbb{P}\big((Q(t) - n)^+ > x\big) \leq \mathbb{P}\bigg(\sup_{0 \leq s \leq t} \big(A(s) - \sum_{i=1}^{n} N_i(s)\big) > x\bigg).$$

If in addition $Q(t)$ converges weakly to a stationary distribution $Q(\infty)$ as $t \to \infty$, then for all $x > 0$,

$$\mathbb{P}\big((Q(\infty) - n)^+ > x\big) \leq \mathbb{P}\bigg(\sup_{t \geq 0} \big(A(t) - \sum_{i=1}^{n} N_i(t)\big) > x\bigg).$$

Note that our bounds are monotone in time, as when t increases the supremum appearing in Theorem 3 is taken over a larger time window, and the bound for the steady-state is the natural limit of these transient bounds.

We will prove Theorem 3 by analyzing a different queueing system $\tilde{\mathcal{Q}}$ which represents a 'modified' queue, in which all servers are kept busy at all times by adding artificial arrivals whenever a server would otherwise go idle. We note that our construction is similar to several constructions appearing in the literature. Our bounding system is closely related to the so-called Queue with Autonomous Service, a model studied previously by several authors [6],[20],[49],[27], whose dynamics can be described as the solution to an appropriate Skorokhod problem [49], [41]. Another related work is [7], in which the queue length of the $G/GI/1$ queue is bounded by considering a modified system in which the server goes on a vacation whenever it would have otherwise gone idle. Also, in [18], the queue length of the $GI/M/n$ queue is bounded by considering a modified system in which a reflecting barrier is placed at state n.

We now construct the FCFS $G/GI/n$ queue $\tilde{\mathcal{Q}}$ on the same probability space as $\{N_i(t), i = 1, \ldots, n\}$ and $A(t)$. We begin by defining two auxiliary processes $\tilde{A}(t)$ and $\tilde{Q}(t)$, where $\tilde{A}(t)$ will become the arrival process to $\tilde{\mathcal{Q}}$, and we will later prove that $\tilde{Q}(t)$ equals the number in system in $\tilde{\mathcal{Q}}$ at time t. Let $\tau_0 \triangleq 0$, $\{\tau_k, k \geq 1\}$ denote the sequence of event times in the pooled renewal process $A(t) + \sum_{i=1}^{n} N_i(t)$, $dA(t) \triangleq A(t) - A(t^-)$, $A(s,t) \triangleq A(t) - A(s)$, and $dN_i(t) \triangleq N_i(t) - N_i(t^-)$, $N_i(s,t) \triangleq N_i(t) - N_i(s)$ for $i = 1, \ldots, n$.

We now define the processes $\tilde{A}(t)$ and $\tilde{Q}(t)$ inductively over $\{\tau_k, k \geq 0\}$. Let $\tilde{A}(\tau_0) \triangleq 0$, $\tilde{Q}(\tau_0) \triangleq n$. Now suppose that for some $k \geq 0$, we have defined $\tilde{A}(t)$ and $\tilde{Q}(t)$ for all $t \leq \tau_k$. We now define these

processes for $t \in (\tau_k, \tau_{k+1}]$. For $t \in (\tau_k, \tau_{k+1})$, let $\tilde{A}(t) \stackrel{\Delta}{=} \tilde{A}(\tau_k)$, and $\tilde{Q}(t) \stackrel{\Delta}{=} \tilde{Q}(\tau_k)$. Note that w.p.1 $dA(\tau_{k+1}) + \sum_{i=1}^{n} dN_i(\tau_{k+1}) = 1$, since $R(A)$ and $R(S)$ are continuous r.v.s, $\mathbb{P}(A = 0) = \mathbb{P}(S = 0) = 0$, and $A(t), \{N_i(t), i = 1, \ldots, n\}$ are mutually independent. We define

$$\tilde{A}(\tau_{k+1}) \stackrel{\Delta}{=} \begin{cases} \tilde{A}(\tau_k) + 1 & \text{if } dA(\tau_{k+1}) = 1; \\ \tilde{A}(\tau_k) + 1 & \text{if } \sum_{i=1}^{n} dN_i(\tau_{k+1}) = 1 \text{ and } \tilde{Q}(\tau_k) \leq n; \\ \tilde{A}(\tau_k) & \text{otherwise (i.e. } \sum_{i=1}^{n} dN_i(\tau_{k+1}) = 1 \text{ and } \tilde{Q}(\tau_k) > n). \end{cases}$$

Similarly, we define

$$\tilde{Q}(\tau_{k+1}) \stackrel{\Delta}{=} \begin{cases} \tilde{Q}(\tau_k) + 1 & \text{if } dA(\tau_{k+1}) = 1; \\ \tilde{Q}(\tau_k) & \text{if } \sum_{i=1}^{n} dN_i(\tau_{k+1}) = 1 \text{ and } \tilde{Q}(\tau_k) \leq n; \\ \tilde{Q}(\tau_k) - 1 & \text{otherwise (i.e. } \sum_{i=1}^{n} dN_i(\tau_{k+1}) = 1 \text{ and } \tilde{Q}(\tau_k) > n). \end{cases}$$

Combining the above completes our inductive definition of $\tilde{A}(t)$ and $\tilde{Q}(t)$. Since w.p.1 $\lim_{k\to\infty} \tau_k = \infty$, it follows that w.p.1 both $\tilde{A}(t)$ and $\tilde{Q}(t)$ are well-defined on $[0, \infty)$. We note that it also follows from our construction that w.p.1 both $\tilde{A}(t)$ and $\tilde{Q}(t)$ are r.c.l.l., and define $d\tilde{A}(t) \stackrel{\Delta}{=} \tilde{A}(t) - \tilde{A}(t^-)$.

We now construct the FCFS $G/GI/n$ queue $\tilde{\mathcal{Q}}$ using the auxiliary process $\tilde{A}(t)$. Let V_i^j denote the length of the jth renewal interval in process $N_i(t), j \geq 1, i = 1, \ldots, n$. Then $\tilde{\mathcal{Q}}$ is defined to be the FCFS $G/GI/n$ queue with arrival process $\tilde{A}(t)$ and processing time distribution S, where the jth job assigned to server i (after time 0) is assigned processing time V_i^{j+1} for $j \geq 1, i = 1, \ldots, n$. The initial conditions for $\tilde{\mathcal{Q}}$ are s.t. for $i = 1, \ldots, n$, there is a single job initially being processed on server i with initial processing time V_i^1, and there are zero jobs waiting in queue.

We now analyze $\tilde{\mathcal{Q}}$, proving that

LEMMA 1. *For $i = 1, \ldots, n$, exactly one job departs from server i at each time $t \in \{\sum_{l=1}^{j} V_i^l, j \geq 1\}$, and there are no other departures from server i. Also, no server ever idles in $\tilde{\mathcal{Q}}$, $\tilde{Q}(t)$ equals the number in system in $\tilde{\mathcal{Q}}$ at time t for all $t \geq 0$, and for all $k \geq 1$,*

$$\tilde{Q}(\tau_k) - n = \max\left(0, \tilde{Q}(\tau_{k-1}) - n + dA(\tau_k) - \sum_{i=1}^{n} dN_i(\tau_k)\right). \tag{2}$$

PROOF. The proof proceeds by induction on $\{\tau_k, k \geq 0\}$, with induction hypothesis that the lemma holds for all $t \leq \tau_k$. The base case $k = 0$ follows from the the initial conditions of $\tilde{\mathcal{Q}}$ and $\tilde{Q}(t)$. Thus assume that the induction hypothesis holds for some fixed $k \geq 0$. We first establish the induction step for the statements about the departure process and non-idling of servers. Let us fix some $i \in \{1, \ldots, n\}$. By the induction hypothesis, server i was non-idling on $[0, \tau_k]$, and the set of departure times from server i on $[0, \tau_k]$ was exactly $\{\sum_{l=1}^{j} V_i^l, j = 1, \ldots, N_i(\tau_k)\}$. We claim that the next departure from server i occurs at time $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l$. Indeed, if $N_i(\tau_k) = 0$, the next departure from server i is the first departure from server i, which occurs at time V_i^1. If instead $N_i(\tau_k) > 0$, then the last departure from server i to occur at or before time τ_k occured at time $\sum_{l=1}^{N_i(\tau_k)} V_i^l$. At that time a new job began processing on server i with processing time $V_i^{N_i(\tau_k)+1}$. This job will depart at time $\sum_{l=1}^{!N_i(\tau_k)+1} V_i^l$, verifying the claim. It follows that no server idles on (τ_k, τ_{k+1}), since $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l \in \{\tau_j, j \geq 1\}$, and thus $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l \geq \tau_{k+1}$. We now treat two cases. First, suppose $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l > \tau_{k+1}$. Then there are no departures from server i on $(\tau_k, \tau_{k+1}]$ and the induction step

follows immediately from the induction hypothesis. Alternatively, suppose $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l = \tau_{k+1}$. In this case the next departure from server i occurs at time τ_{k+1}, $dN_i(\tau_{k+1}) = 1$, and all other servers are non-idling and have no departures on $(\tau_k, \tau_{k+1}]$. Thus if there are at least $n+1$ jobs in $\tilde{\mathcal{Q}}$ at time τ_k, then there are at least $n+1$ jobs in $\tilde{\mathcal{Q}}$ at time τ_{k+1}^-, and some job begins processing on server i at time τ_{k+1}. Alternatively, if there are exactly n jobs in $\tilde{\mathcal{Q}}$ at time τ_k, then $\tilde{Q}(\tau_k) = n$ by the induction hypothesis. Thus $d\tilde{A}(\tau_{k+1}) = 1$, and this arrival immediately begins processing on server i. Combining the above treats all cases since there are at least n jobs in $\tilde{\mathcal{Q}}$ at time τ_k by the induction hypothesis, completing the induction step.

We now prove the induction step for the statement that $\tilde{Q}(t)$ equals the number in system in $\tilde{\mathcal{Q}}$ at time t, as well as (2). Since we have already proven that any departures from $\tilde{\mathcal{Q}}$ on $(\tau_k, \tau_{k+1}]$ occur at time τ_{k+1}, and by construction any jumps in $\tilde{A}(t)$ and $\tilde{Q}(t)$ on $(\tau_k, \tau_{k+1}]$ occur at time τ_{k+1}, it suffices to prove that $\tilde{Q}(\tau_{k+1})$ equals the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1}. First, suppose $dA(\tau_{k+1}) = 1$. Then $\sum_{i=1}^n dN_i(\tau_{k+1}) = 0$, $\tilde{Q}(\tau_k) \geq n$ by the induction hypothesis, and $\tilde{Q}(\tau_{k+1}) = \tilde{Q}(\tau_k) + 1$. Thus

$$\begin{aligned} \max\left(0, \tilde{Q}(\tau_k) - n + dA(\tau_{k+1}) - \textstyle\sum_{i=1}^n dN_i(\tau_{k+1})\right) &= \max\left(0, \tilde{Q}(\tau_k) - n + 1\right) \\ &= \tilde{Q}(\tau_k) - n + 1 \\ &= \tilde{Q}(\tau_{k+1}) - n, \end{aligned}$$

showing that (2) holds. Note that $\sum_{i=1}^n dN_i(\tau_{k+1}) = 0$ implies that $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l > \tau_{k+1}$ for all $i = 1, \ldots, n$, and we have already proven that in this case there are no departures from $\tilde{\mathcal{Q}}$ on $(\tau_k, \tau_{k+1}]$. Since $dA(\tau_{k+1}) = 1$ implies $d\tilde{A}(\tau_{k+1}) = 1$, it follows that the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1} is one more than the number in system in $\tilde{\mathcal{Q}}$ at time τ_k. Thus $\tilde{Q}(\tau_{k+1})$ equals the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1} by the induction hypothesis.

Now suppose that $\sum_{i=1}^n dN_i(\tau_{k+1}) = 1$. Then $dA(\tau_{k+1}) = 0$, and there exists a unique index i^* s.t. $\sum_{l=1}^{N_{i^*}(\tau_k)+1} V_{i^*}^l = \tau_{k+1}$. We have already proven that in this case there are no departures from $\tilde{\mathcal{Q}}$ on (τ_k, τ_{k+1}), and a single departure from $\tilde{\mathcal{Q}}$ at time τ_{k+1} (on server i^*). First suppose that there are at least $n+1$ jobs in $\tilde{\mathcal{Q}}$ at time τ_k. Then $\tilde{Q}(\tau_k) \geq n+1$ by the induction hypothesis, and $\tilde{Q}(\tau_{k+1}) = \tilde{Q}(\tau_k) - 1$. Thus

$$\begin{aligned} \max\left(0, \tilde{Q}(\tau_k) - n + dA(\tau_{k+1}) - \textstyle\sum_{i=1}^n dN_i(\tau_{k+1})\right) &= \max\left(0, \tilde{Q}(\tau_k) - n - 1\right) \\ &= \tilde{Q}(\tau_k) - n - 1 \\ &= \tilde{Q}(\tau_{k+1}) - n, \end{aligned}$$

showing that (2) holds. Since $d\tilde{A}(\tau_{k+1}) = 0$, there are no arrivals to $\tilde{\mathcal{Q}}$ on $(\tau_k, \tau_{k+1}]$. Combining the above, we find that the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1} is one less than the number in system in $\tilde{\mathcal{Q}}$ at time τ_k. Thus $\tilde{Q}(\tau_{k+1})$ equals the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1} by the induction hypothesis.

Alternatively, suppose that $\sum_{i=1}^n dN_i(\tau_{k+1}) = 1$ and there are exactly n jobs in $\tilde{\mathcal{Q}}$ at time τ_k. Then $\tilde{Q}(\tau_k) = n$ by the induction hypothesis, and $\tilde{Q}(\tau_{k+1}) = \tilde{Q}(\tau_k)$. Thus

$$\begin{aligned} \max\left(0, \tilde{Q}(\tau_k) - n + dA(\tau_{k+1}) - \textstyle\sum_{i=1}^n dN_i(\tau_{k+1})\right) &= \max\left(0, \tilde{Q}(\tau_k) - n - 1\right) \\ &= 0 \\ &= \tilde{Q}(\tau_{k+1}) - n, \end{aligned}$$

showing that (2) holds. Since $d\tilde{A}(\tau_{k+1}) = 1$, there is a single arrival to $\tilde{\mathcal{Q}}$ on $(\tau_k, \tau_{k+1}]$. Combining the above, we find that the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1} equals the number in system

in $\tilde{\mathcal{Q}}$ at time τ_k. Thus $\tilde{Q}(\tau_{k+1})$ equals the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1} by the induction hypothesis. Since $\tilde{Q}(\tau_k) \geq n$ by the induction hypothesis, this treats all cases, completing the proof of the induction and the lemma. □

We now 'unfold' recursion (2) to derive a simple one-dimensional random walk representation for $\tilde{Q}(t)$. The relationship between recursions such as (2) and the suprema of associated one-dimensional random walks is well-known (see [6],[7]), and can also be formalized by studying the appropriate Skorokhod problem [41]. Furthermore, although it seems that $\tilde{Q}(t) - n$ cannot be immediately related to the Skorokhod problem naturally associated with $Q(t) - n$, we note that such a formulation may be possible through the framework of jump reflection (see [31]).

Then it follows from (2) and a straightforward induction on $\{\tau_k, k \geq 0\}$ that w.p.1, for all $k \geq 0$,

$$\tilde{Q}(\tau_k) - n = \max_{0 \leq j \leq k} \left(A\big(\tau_{k-j}, \tau_k\big) - \sum_{i=1}^{n} N_i\big(\tau_{k-j}, \tau_k\big) \right).$$

As all jumps in $\tilde{Q}(t)$ occur at times $t \in \{\tau_k, k \geq 1\}$, it follows that

COROLLARY 1. *W.p.1, for all $t \geq 0$,*

$$\tilde{Q}(t) - n = \sup_{0 \leq s \leq t} \left(A\big(t-s, t\big) - \sum_{i=1}^{n} N_i\big(t-s, t\big) \right).$$

We now prove that $\tilde{Q}(t)$ provides an upper bound for $Q(t)$.

PROPOSITION 1. *$Q(t)$ and $\tilde{Q}(t)$ can be constructed on the same probability space so that w.p.1 $Q(t) \leq \tilde{Q}(t)$ for all $t \geq 0$.*

For our later results, it will be useful to first prove a general comparison result for $G/G/n$ queues. Although such results seem to be generally known in the queueing literature (see [46],[40]), we include a proof for completeness. For an event E, let $I(E)$ denote the indicator function of E.

LEMMA 2. *Let $\mathcal{Q}^1$ and $\mathcal{Q}^2$ be two FCFS $G/G/n$ queues with finite, strictly positive inter-arrival and processing times. Let $\{T_k^i, k \geq 1\}$ denote the ordered sequence of arrival times to $\mathcal{Q}^i$, $i \in \{1, 2\}$. Let S_k^i denote the processing time assigned to the job that arrives to $\mathcal{Q}^i$ at time T_k^i, $k \geq 1, i \in \{1, 2\}$. Further suppose that*

(i) The initial number in system in $\mathcal{Q}^1$ is at most n;
(ii) For each job J initially in $\mathcal{Q}^1$ there is a distinct corresponding job J' initially in $\mathcal{Q}^2$ s.t. the initial processing time of J in $\mathcal{Q}^1$ equals the initial processing time of J' in $\mathcal{Q}^2$;
(iii) $\{T_k^1, k \geq 1\}$ is a subsequence of $\{T_k^2, k \geq 1\}$;
(iv) For all $k \geq 1$, the job that arrives to $\mathcal{Q}^2$ at time T_k^1 is assigned processing time S_k^1, the same processing time assigned to the job which arrives to $\mathcal{Q}^1$ at that time.

Then the number in system in $\mathcal{Q}^2$ at time t is at least the number in system in $\mathcal{Q}^1$ at time t for all $t \geq 0$.

PROOF. The proof is deferred to the appendix. □

We now complete the proof of Proposition 1.

PROOF OF PROPOSITION 1. We construct $\tilde{\mathcal{Q}}$ and $\mathcal{Q}$ on the same probability space. We assign $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$ the same initial conditions, and let $A(t)$ be the arrival process to $\mathcal{Q}$ on $(0,\infty)$. Let $\{t_k, k \geq 1\}$ denote the ordered sequence of event times in $A(t)$. It follows from the construction of $\tilde{A}(t)$ that $\{t_k, k \geq 1\}$ is a subsequence of the set of event times in $\tilde{A}(t)$. We let the processing time assigned to the arrival to $\tilde{\mathcal{Q}}$ at time t_k equal the processing time assigned to the arrival to $\mathcal{Q}$ at time t_k, $k \geq 1$. It follows that w.p.1 $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$ satisfy the conditions of Lemma 2. Combining the above with Lemma 1 completes the proof. □

We now complete the proof of Theorem 3.

PROOF OF THEOREM 3. By elementary renewal theory (see [9]), $A(s)_{0\leq s\leq t}$ has the same distribution (on the process level) as $A\big(t-s,t\big)_{0\leq s\leq t}$, and $\sum_{i=1}^n N_i(s)_{0\leq s\leq t}$ has the same distribution (on the process level) as $\sum_{i=1}^n N_i\big(t-s,t\big)_{0\leq s\leq t}$. Combining with the independence of $A(t)$ and $\sum_{i=1}^n N_i(t)$, Corollary 1, and Proposition 1, proves the theorem.

We now prove the corresponding steady-state result. Note that for any $x > 0$, the sequence of events $\big\{\sup_{0\leq s\leq t}\big(A(s)-\sum_{i=1}^n N_i(s)\big) > x, t \geq 0\big\}$ is monotonic in t. It follows from the continuity of probability measures that

$$\lim_{t\to\infty}\mathbb{P}\bigg(\sup_{0\leq s\leq t}\big(A(s)-\sum_{i=1}^n N_i(s)\big) > x\bigg) = \mathbb{P}\bigg(\sup_{t\geq 0}\big(A(t)-\sum_{i=1}^n N_i(t)\big) > x\bigg).$$

The steady-state result then follows from the corresponding transient result and the definition of weak convergence, since $Q(\infty)$ has integer support. □

4. Lower bound. In this section, we prove a general lower bound for the $M/GI/n$ queue, when properly initialized. Suppose A is an exponentially distributed r.v. Let Z denote a Poisson r.v. with mean $\frac{\mu_A}{\mu_S}$. Let $\mathcal{Q}_2$ denote the $M/GI/n$ queue with inter-arrival times drawn i.i.d. distributed as A, processing times drawn i.i.d. distributed as S, and the following initial conditions. At time 0 there are Z jobs in system. This set of initial jobs have initial processing times drawn i.i.d. distributed as $R(S)$, independent of Z. If $Z \geq n$, a set of exactly n initial jobs is selected uniformly at random (u.a.r.) to be processed initially, and the remaining initial jobs queue for processing. Suppose also that the first inter-arrival time is distributed as $R(A)$ (also an exponentially distributed r.v.) independent of both Z and the initial processing times of those jobs initially in the system. Recall the processes $A(t)$ and $\{N_i(t)!, i = 1,\ldots,n\}$, which were defined previously at the start of Section 3. Then $Q_2(t)$, the number in system at time t in $\mathcal{Q}_2$, satisfies

THEOREM 4. *For all $x > 0$, and $t \geq 0$,*

$$\mathbb{P}\big((Q_2(t)-n)^+ > x\big) \geq \mathbb{P}\big(Z \geq n\big)\sup_{0\leq s\leq t}\mathbb{P}\big(A(s)-\sum_{i=1}^n N_i(s) > x\big).$$

If in addition $Q_2(t)$ converges weakly to a stationary distribution $Q(\infty)$ as $t \to \infty$, then for all $x > 0$,

$$\mathbb{P}\big((Q(\infty)-n)^+ > x\big) \geq \mathbb{P}\big(Z \geq n\big)\sup_{t\geq 0}\mathbb{P}\big(A(t)-\sum_{i=1}^n N_i(t) > x\big).$$

Comparing with Theorem 3, we see that our upper and lower bounds exhibit a certain duality, marked by the order of the $\mathbb{P}$ and sup operators.

We will prove Theorem 4 by coupling $\mathcal{Q}_2$ to *both* an associated FCFS $M/GI/\infty$ queue $\mathcal{Q}_\infty$ and a certain family of FCFS $G/G/n$ queues $\{\mathcal{Q}_2^s, s \geq 0\}$. For each $s \geq 0$, our coupling ensures that $Q_2^s(t)$, the number in system at time t in $\mathcal{Q}_2^s$, provides a lower bound for $Q_2(t)$ for all $t \geq s$, and that the set of remaining processing times (at time s) of those jobs in $\mathcal{Q}_2^s$ at time s is a random thinning of the set of remaining processing times (at time s) of those jobs in $\mathcal{Q}_\infty$ at time s. We note that some of the ideas involved in the proof of our lower bound have appeared in the literature before (see [43], [40], [48]).

We now construct $\mathcal{Q}_\infty$ and $\{\mathcal{Q}_2^s, s \geq 0\}$. We assign $\mathcal{Q}_\infty$ the same initial conditions as $\mathcal{Q}_2$ (although in $\mathcal{Q}_\infty$ all initial jobs begin processing at time 0). We let $\mathcal{Q}_\infty$ and $\mathcal{Q}_2$ have the same arrival process, and for each arrival, we let the processing time assigned to this arrival to $\mathcal{Q}_\infty$ equal the processing time assigned to this arrival to $\mathcal{Q}_2$.

We now describe the initial conditions and arrival process for $\mathcal{Q}_2^s$ in terms of an appropriate thinning of the initial conditions and arrival process of $\mathcal{Q}_\infty$, where the nature of this thinning depends on $Q_\infty(s)$, the number in system at time s in $\mathcal{Q}_\infty$. If $Q_\infty(s) < n$, then the initial conditions of $\mathcal{Q}_2^s$ are to have zero jobs in system, and the arrival process to $\mathcal{Q}_2^s$ is to have zero arrivals on $[0, \infty)$. If $Q_\infty(s) \geq n$, then we select a size-n subset $\mathcal{C}^s$ of jobs u.a.r. from all subsets of the jobs being processed in $\mathcal{Q}_\infty$ at time s. Let $\mathcal{C}_0^s$ denote those jobs in $\mathcal{C}^s$ which were initially in $\mathcal{Q}_\infty$ at time 0. Then the initial conditions of $\mathcal{Q}_2^s$ are as follows. For each job $J \in \mathcal{C}_0^s$, there is a corresponding job J' initially in $\mathcal{Q}_2^s$, where the initial processing time of J' in $\mathcal{Q}_2^s$ equals the initial processing time of J in $\mathcal{Q}_\infty$. There are no other initial jobs in $\mathcal{Q}_2^s$. The arrival process to $\mathcal{Q}_2^s$ on $(0, s]$ is as follows. For each job J that arrives to $\mathcal{Q}_\infty$ (and thus to $\mathcal{Q}_2$) on $(0, s]$, say at time τ, there is a corresponding arrival J' to $\mathcal{Q}_2^s$ at time τ iff $J \in \mathcal{C}^s \setminus \mathcal{C}_0^s$. In this case, the processing time assigned to J' in $\mathcal{Q}_2^s$ equals the processing time assigned to J in $\mathcal{Q}_\infty$. There are no other arrivals to $\mathcal{Q}_2^s$ on $(0, s]$. We let $\mathcal{Q}_2^s$, $\mathcal{Q}_\infty$, and $\mathcal{Q}_2$ have the same arrival process on (s, ∞), and for each arrival, we let the processing time assigned to this arrival to $\mathcal{Q}_2^s$ equal the processing time assigned to this arrival to $\mathcal{Q}_\infty$(and thus $\mathcal{Q}_2$).

We claim that our coupling of $\mathcal{Q}_\infty$ to $\mathcal{Q}_2$ and construction of $\mathcal{Q}_2^s$ ensure that $\mathcal{Q}_2^s$ and $\mathcal{Q}_2$ satisfy the conditions of Lemma 2. Indeed, for each job initially in $\mathcal{Q}_2^s$, there is a distinct corresponding job initially in $\mathcal{Q}_2$ with the same initial processing time. Also, for each job that arrives to $\mathcal{Q}_2^s$, there is a distinct corresponding job that arrives to $\mathcal{Q}_2$ at the same time with the same processing time. Thus w.p.1 $Q_2^s(t)$, the number in system at time t in $\mathcal{Q}_2^s$, satisfies

$$Q_2(t) \geq Q_2^s(t) \text{ for all } s, t \geq 0. \tag{3}$$

We now complete the proof of Theorem 4.

Proof of Theorem 4. Since $\mathcal{Q}_\infty$ is initialized with its stationary measure (see [45]), it follows from the basic properties of the $M/GI/\infty$ queue (see [45]) that $\mathbb{P}(Q_\infty(s) \geq n) = \mathbb{P}(Z \geq n)$, and conditional on the event $\{Q_\infty(s) \geq n\}$, the set of remaining processing times (at time s) of those jobs being processed in $\mathcal{Q}_\infty$ at time s are drawn i.i.d. distributed as $R(S)$. Thus conditional on the event $\{Q_\infty(s) \geq n\}$, one has that $|\mathcal{C}^s| = n$, and the set of remaining processing times (at time s, in $\mathcal{Q}_\infty$) of those jobs belonging to $\mathcal{C}^s$ is drawn i.i.d. distributed as $R(S)$.

By construction the number of jobs initially in $\mathcal{Q}_2^s$ at time 0 *plus* the number of jobs that arrive to $\mathcal{Q}_2^s$ on $(0, s]$ is at most n. Thus all jobs initially in $\mathcal{Q}_2^s$ at time 0 and all jobs that arrive to $\mathcal{Q}_2^s$ on $(0, s]$ begin processing immediately in $\mathcal{Q}_2^s$, as if $\mathcal{Q}_2^s$ were an infinite-server queue. It follows from our construction that conditional on the event $\{Q_\infty(s) \geq n\}$, the set of remaining processing times

(at time s) of the n jobs in $\mathcal{Q}_2^s$ at time s equals the set of remaining processing times (at time s, in $\mathcal{Q}_\infty$) of those jobs belonging to $\mathcal{C}^s$, and are thus drawn i.i.d. distributed as $R(S)$.

Let us fix some s, t s.t. $0 \leq s \leq t$. Recall that V_i^j denotes the length of the jth renewal interval in process $N_i(t), j \geq 1, i = 1, \ldots, n$. It follows from our construction that conditional on the event $\{Q_\infty(s) \geq n\}$, we may set the remaining processing time (at time s) of the job on server i in $\mathcal{Q}_2^s$ at time s equal to V_i^1. We can also set the processing time of the jth job assigned to server i in $\mathcal{Q}_2^s$ (after time s) equal to V_i^{j+1}. Under this coupling the total number of jobs that depart from server i in $\mathcal{Q}_2^s$ during $[s,t]$ is at most $N_i(t-s)$, and therefore the total number of departures from $\mathcal{Q}_2^s$ during $[s,t]$ is at most $\sum_{i=1}^n N_i(t-s)$, independent of the arrival process to $\mathcal{Q}_2^s$ on $[s,t]$. By the memoryless and stationary increments properties of the Poisson process, we may let the arrival process to $\mathcal{Q}_2^s$ on $[s,t]$ equal $A(v)_{0 \leq v \leq t-s}$. Combining the above, we find that for all $x > 0$, $\mathbb{P}(Q_2^s(t) - n > x) \geq \mathbb{P}(Z \geq n)\mathbb{P}\big(A(t-s) - \sum_{i=1}^n N_i(t-s) > x\big)$. Observing that s was general, we may then take the supremum of the above bound over all $s \in [0,t]$, and combine with (3) to complete the proof of the theorem. The corresponding steady-state result then follows from the fact that monotonic sequences have limits and the definition of weak convergence. $\square$

5. Tightness and proof of Theorem 1. In this section, we prove Theorem 1. We note that it follows almost immediately from Theorem 3 and well-known tightness results from the literature (see [5] Theorem 14.6, [49] Theorem 7.2.3) that for any *fixed* $T \geq 0$, $\{n^{-\frac{1}{2}}\big(Q^n(t) - n\big)^+_{0 \leq t \leq T}, n \geq 1\}$ is tight in the space $D[0,T]$ under the J_1 topology (see Subsection 6.1 for details). The challenge is that when analyzing $\{n^{-\frac{1}{2}}\big(Q^n(\infty) - n\big)^+, n \geq 1\}$, one does not have the luxury of bounded time intervals. In particular, to apply Theorem 3, we must show tightness of a supremum taken over an infinite time horizon. For this reason, most standard weak convergence type results and arguments from the literature (see [49]) break down, and cannot immediately be applied. Instead, we will relate the supremum appearing in the right hand side (r.h.s.) of Theorem 3 to the steady-state waiting time in an appropriate $G/D/1$ queue with stationary (as opposed to i.i.d.) inter-arrival times. We will then apply known results from the literature, in particular [44], to show that under the H-W scaling this sequence of steady-state waiting times, properly normalized, is tight.

Suppose that assumptions H-W and T_0 hold. Let $A_n(t) \stackrel{\Delta}{=} A(\lambda_n t)$. In light of Theorem 3, it suffices to prove that $\{n^{-\frac{1}{2}} \sup_{t \geq 0}\big(A_n(t) - \sum_{i=1}^n N_i(t)\big), n \geq 1\}$ is tight. Let $\mathsf{A}_n^0(t)$ denote an ordinary renewal process with renewal distribution $A\lambda_n^{-1}$, independent of $\{N_i(t), i = 1, \ldots, n\}$. Note that we may construct $A_n(t)$ and $A_n^0(t)$ on the same probability space so that $A_n(t) \leq 1 + A_n^0(t)$ for all $t \geq 0$. It thus suffices to demonstrate the tightness of $\{n^{-\frac{1}{2}} \sup_{t \geq 0}\big(A_n^0(t) - \sum_{i=1}^n N_i(t)\big), n \geq 1\}$.

Let $\{A_i^1, i \geq 1\}$ denote a countably infinite sequence of r.v.s drawn i.i.d. distributed as A, independent of $\{N_i(t), i = 1, \ldots, n\}$. Note that since $A_n^0(t) - \sum_{i=1}^n N_i(t)$ only increases at jumps of $A_n^0(t)$, we may construct $A_n^0(t), \sum_{i=1}^n N_i(t)$, and $\{A_i^1, i \geq 1\}$ on the same probability space so that

$$n^{-\frac{1}{2}} \sup_{t \geq 0}\big(A_n^0(t) - \sum_{i=1}^n N_i(t)\big) = n^{-\frac{1}{2}} \sup_{k \geq 0}\big(k - \sum_{i=1}^n N_i(\lambda_n^{-1} \sum_{j=1}^k A_j^1)\big). \tag{4}$$

We now show that

$$\big\{n^{-\frac{1}{2}} \sup_{k \geq 0}\big(k - \sum_{i=1}^n N_i(\lambda_n^{-1} \sum_{j=1}^k A_j^1)\big), n \geq 1\big\} \tag{5}$$

is tight, which (by the above) will imply Theorem 1. Fortunately, the tightness of such sequences of suprema has already been addressed in the literature, in the context of steady-state waiting

times in a $G/G/1$ queue, with stationary inter-arrival times, in heavy-traffic. In particular, note that for $M \geq 1$, $\sup_{0\leq k\leq M}\big(k - \sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1)\big)$ corresponds to the waiting time of the $(M+1)$st arrival to a $G/D/1$ queue, initially empty, with all processing times equal to 1, and the kth inter-arrival time equal to

$$\sum_{i=1}^{n} N_i\big(\lambda_n^{-1}\sum_{j=1}^{M-k} A_j^1, \lambda_n^{-1}\sum_{j=1}^{M-k+1} A_j^1\big), k \leq M.$$

Recall that $\sum_{i=1}^{n} N_i(t)_{t\geq 0}$ has the same distribution (on the process level) as $\sum_{i=1}^{n} N_i\big(t-s,t\big)_{0\leq s\leq t}$ (see [9]), and $\{A_i^1, i \geq 1\}$ are i.i.d. It follows that for all $M \geq 1$, $\sup_{0\leq k\leq M}\big(k - \sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1)\big)$ also has the same distribution as the waiting time of the $(M+1)$st arrival to a $G/D/1$ queue, initially empty, with all processing times equal to 1, and the kth inter-arrival time equal to

$$\sum_{i=1}^{n} N_i\big(\lambda_n^{-1}\sum_{j=1}^{k-1} A_j^1, \lambda_n^{-1}\sum_{j=1}^{k} A_j^1\big), k \geq 1.$$

For this queueing model, in which the sequence of inter-arrival times is stationary, one can ask whether there is a meaningful notion of steady-state waiting time, whose distribution would naturally coincide with that of

$$\lim_{M\to\infty}\sup_{0\leq k\leq M}\big(k - \sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1)\big) = \sup_{k\geq 0}\big(k - \sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1)\big).$$

Furthermore, should one examine a sequence of such queues in heavy traffic, one can ask whether the corresponding sequence of steady-state waiting times, properly normalized, is tight.

Note that as (5) is such a sequence, we are left to answer exactly this question. Fortunately, sufficient conditions for tightness of such a sequence are given in [44]. In particular, as we will show, it follows from the results of [44] (in the notation of [44]) that

THEOREM 5. *Suppose that for all sufficiently large n, $\{\zeta_{n,i}, i \geq 1\}$ is a stationary, countably infinite sequence of r.v. Let $a_n \stackrel{\Delta}{=} \mathbb{E}[\zeta_{n,1}]$, and $W_{n,k} \stackrel{\Delta}{=} \sum_{i=1}^{k}\zeta_{n,i}$. Further assume that $a_n < 0, \lim_{n\to\infty} a_n = 0$, and there exist $C_1, C_2 < \infty$ and $\epsilon > 0$ s.t. for all sufficiently large n,*

(i) $\mathbb{E}\big[|W_{n,k} - ka_n|^{2+\epsilon}\big] \leq C_1 k^{1+\frac{\epsilon}{2}}$ *for all* $k \geq 1$;
(ii) $\mathbb{P}\big(\max_{i=1,\ldots,k}(W_{n,i} - ia_n) > x\big) \leq C_2 k^{1+\frac{\epsilon}{2}} x^{-(2+\epsilon)}$ *for all* $k \geq 1$ *and* $x > 0$.

Then $\{|a_n|\sup_{k\geq 0} W_{n,k}, n \geq 1\}$ *is tight.*

PROOF. The proof follows from Theorem 1 of [44], and is deferred to the appendix. □

To verify that the assumptions of Theorem 5 hold for

$$\big\{n^{-\frac{1}{2}}\sup_{k\geq 0}\big(k - \sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1)\big), n \geq 1\big\},$$

we will rely on a technical result from [5], which gives a bound on the supremum of a general random walk in terms of bounds on its increments. In particular, it is shown in [5] Theorem 10.2 that

LEMMA 3. *Suppose $k < \infty$, $X_1, X_2, \ldots, X_k$ is a sequence of general (possibly dependent and not identically distributed) random variables, $S_j \stackrel{\Delta}{=} \sum_{i=1}^{j} X_i$, and $M_k = \max_{j \le k} |S_j|$. Further suppose that there exist real numbers $\alpha > \frac{1}{2}$, $\beta \ge 0$, and a sequence of non-negative numbers $u_1, u_2, \ldots, u_k$ s.t. for all $0 \le i \le j \le k$ and $x > 0$,*

$$\mathbb{P}\big(|S_j - S_i| \ge x\big) \le x^{-4\beta}\big(\sum_{i<l\le j} u_l\big)^{2\alpha}.$$

Then there exists a finite constant $K_{\alpha,\beta}$, depending only on α and β, s.t. for all $x > 0$,

$$\mathbb{P}\big(M_k \ge x\big) \le K_{\alpha,\beta} x^{-4\beta}\big(\sum_{0<l\le k} u_l\big)^{2\alpha}.$$

We will also use frequently the inequality

$$(x_1 + x_2)^r \le 2^{r-1}x_1^r + 2^{r-1}x_2^r \text{ for all } r \ge 1 \text{ and } x_1, x_2 \ge 0, \tag{6}$$

which follows from the convexity of $f(x) \stackrel{\Delta}{=} x^r$, $r \ge 1$.

Before proceeding with the proof of Theorem 1, we establish two more auxiliary results. The first bounds the moments of the sum of n i.i.d. zero-mean r.v. in terms of the moments of the individual r.v.s and n, and is proven in [50].

LEMMA 4. *For all $r \ge 2$, there exists $C_r < \infty$ (depending only on r) s.t. for all r.v. X satisfying $\mathbb{E}[X] = 0$ and $\mathbb{E}[|X|^r] < \infty$, if $\{X_i, i \ge 1\}$ is a sequence of i.i.d. r.v.s distributed as X, then for all $k \ge 1$,*

$$\mathbb{E}[|\sum_{i=1}^{k} X_i|^r] \le C_r k^{\frac{r}{2}} \mathbb{E}[|X|^r].$$

Second, we prove a bound for the central moments of a pooled equilibrium renewal process.

LEMMA 5. *Let X denote any non-negative r.v. s.t. $\mathbb{E}[X] = \mu^{-1} \in (0, \infty)$, and $\mathbb{E}[X^r] < \infty$ for some $r \ge 2$. Let $\{Z_i^e(t), i \ge 1\}$ denote a set of i.i.d. equilibrium renewal processes with renewal distribution X. Then there exists $C_{X,r} < \infty$ (depending only on X and r) s.t. for all $n \ge 1$ and $t \ge 0$,*

$$\mathbb{E}\big[|\sum_{i=1}^{n} Z_i^e(t) - \mu n t|^r\big] \le C_{X,r}\big(1 + (nt)^{\frac{r}{2}}\big). \tag{7}$$

PROOF. The proof is deferred to the appendix. □

With the above bounds at our disposal, we now complete the proof of Theorem 1.

PROOF OF THEOREM 1. In the notation of Theorem 5, let

$$\zeta_{n,k} \stackrel{\Delta}{=} 1 - \sum_{i=1}^{n} N_i(\lambda_n^{-1} \sum_{j=1}^{k-1} A_j^1, \lambda_n^{-1} \sum_{j=1}^{k} A_j^1),$$

$$W_{n,k} \stackrel{\Delta}{=} k - \sum_{i=1}^{n} N_i(\lambda_n^{-1} \sum_{j=1}^{k} A_j^1).$$

That $\{\zeta_{n,i}, i \geq 1\}$ is a stationary, countably infinite sequence of r.v. follows from the stationary increments property of the equilibrium renewal process. Since $\mathbb{E}[\sum_{i=1}^{n} N_i(t)] = nt\mu$ for all $t \geq 0$, it follows that $a_n \stackrel{\Delta}{=} \mathbb{E}[\zeta_{n,1}] = 1 - \frac{n}{\lambda_n} = -\frac{B}{n^{\frac{1}{2}} - B} < 0$, and $\lim_{n\to\infty} a_n = 0$. Thus we need only verify assumptions (i) and (ii) of Theorem 5. Since $\mathbb{E}[A^{2+\epsilon}], \mathbb{E}[S^{2+\epsilon}] < \infty$ for some $\epsilon > 0$ by the T_0 assumptions, we may fix some $r > 2$ s.t. $\mathbb{E}[A^r], \mathbb{E}[S^r] < \infty$. Note that

$$\begin{aligned}
\mathbb{E}\big[|W_{n,k} - ka_n|^r\big] &= \mathbb{E}\big[|\sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1) - \frac{kn}{\lambda_n}|^r\big] \\
&\leq \mathbb{E}\big[\Big(|\sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1) - \mu\frac{n}{\lambda_n}\sum_{j=1}^{k} A_j^1| + |\mu\frac{n}{\lambda_n}\sum_{j=1}^{k} A_j^1 - \frac{kn}{\lambda_n}|\Big)^r\big] \quad \text{by the tri. ineq.} \\
&\leq 2^{r-1}\mathbb{E}\big[|\sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1) - \mu\frac{n}{\lambda_n}\sum_{j=1}^{k} A_j^1|^r\big] \qquad (8) \\
&\quad + 2^{r-1}\mathbb{E}\big[|\mu\frac{n}{\lambda_n}\sum_{j=1}^{k} A_j^1 - \frac{kn}{\lambda_n}|^r\big] \quad \text{by (6).} \qquad (9)
\end{aligned}$$

We now bound (8). By Lemmas 4 - 5, there exist $C_{S,r}, C_r < \infty$ independent of n and k s.t.

$$\begin{aligned}
\mathbb{E}\big[|\sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1) - \mu\frac{n}{\lambda_n}\sum_{j=1}^{k} A_j^1|^r\big] &\leq C_{S,r} + C_{S,r}(\frac{n}{\lambda_n})^{\frac{r}{2}}\mathbb{E}\big[(\sum_{j=1}^{k} A_j^1)^{\frac{r}{2}}\big] \text{ by Lemma 5} \\
&\leq C_{S,r} + C_{S,r}(\frac{n}{\lambda_n})^{\frac{r}{2}}\mathbb{E}\big[\Big(|\sum_{j=1}^{k}\big(A_j^1 - \mu^{-1}\big)| + k\mu^{-1}\Big)^{\frac{r}{2}}\big] \\
&\leq C_{S,r} + C_{S,r}(\frac{n}{\lambda_n})^{\frac{r}{2}}\Big(2^{\frac{r}{2}-1}\mathbb{E}\big[|\sum_{j=1}^{k}\big(A_j^1 - \mu^{-1}\big)|^{\frac{r}{2}}\big] + 2^{\frac{r}{2}-1}(k\mu^{-1})^{\frac{r}{2}}\Big) \\
&\leq C_{S,r} + 2^{\frac{r}{2}-1}C_{S,r}(\frac{n}{\lambda_n})^{\frac{r}{2}}\Big(\mathbb{E}^{\frac{1}{2}}\big[|\sum_{j=1}^{k}(A_j^1 - \mu^{-1})|^r\big] + (k\mu^{-1})^{\frac{r}{2}}\Big) \\
&\quad \text{since } \mathbb{E}[X] \leq \mathbb{E}^{\frac{1}{2}}[X^2] \text{ for any non-negative r.v. } X \\
&\leq C_{S,r} + 2^{\frac{r}{2}-1}C_{S,r}(\frac{n}{\lambda_n})^{\frac{r}{2}}\Big(\big(C_r k^{\frac{r}{2}}\mathbb{E}\big[|A - \mu^{-1}|^r\big]\big)^{\frac{1}{2}} + (k\mu^{-1})^{\frac{r}{2}}\Big) \\
&\quad \text{by Lemma 4.} \\
&\leq C_1' k^{\frac{r}{2}} \qquad (10)
\end{aligned}$$

for some finite constant C_1' independent of n and k, since $\mathbb{E}\big[|A - \mu^{-1}|^r\big] < \infty$, and $\lim_{n\to\infty} \frac{n}{\lambda_n} = 1$.

We now bound (9).

$$\begin{aligned}
\mathbb{E}\big[|\mu\frac{n}{\lambda_n}\sum_{j=1}^{k} A_j^1 - \frac{kn}{\lambda_n}|^r\big] &= (\frac{n}{\lambda_n})^r\mu^r\mathbb{E}\big[|\sum_{j=1}^{k}(A_j^1 - \mu^{-1})|^r\big] \\
&\leq \Big(C_r(\frac{n}{\lambda_n})^r\mu^r\mathbb{E}\big[|A - \mu^{-1}|^r\big]\Big)k^{\frac{r}{2}} \text{ by Lemma 4} \\
&\leq C_1'' k^{\frac{r}{2}} \quad \text{for some finite constant } C_1'' \text{ independent of } n \text{ and } k. \qquad (11)
\end{aligned}$$

Using (10) to bound (8) and (11) to bound (9), it follows that assumption (i) of Theorem 5 holds for the finite constant $C_1 \stackrel{\Delta}{=} 2^{r-1}(C_1' + C_1'')$. We now apply Lemma 3 to show that assumption (ii) holds as well. In the notation of Lemma 3, let $S_{n,i} \stackrel{\Delta}{=} W_{n,i} - ia_n$ for $i \geq 0$, and $M_{n,k} \stackrel{\Delta}{=} \max_{i \leq k} |W_{n,i} - ia_n|$ for $k \geq 0$. Then for all n, $0 \leq i \leq j$, and $x > 0$,

$$\begin{aligned}
\mathbb{P}\big(|S_{n,j} - S_{n,i}| \geq x\big) &= \mathbb{P}\big(|S_{n,j-i}| \geq x\big) \text{ by stationary increments} \\
&= \mathbb{P}(|W_{n,j-i} - (j-i)a_n| \geq x) \\
&\leq C_1(j-i)^{\frac{r}{2}} x^{-r} \text{ by Markov's inequality} \\
&\leq \big((C_1+1)(j-i)\big)^{\frac{r}{2}} x^{-r}.
\end{aligned}$$

Thus for all n and $k \geq 1$, we may apply Lemma 3 (in the notation of Lemma 3) with $\beta \stackrel{\Delta}{=} \frac{r}{4}, \alpha \stackrel{\Delta}{=} \frac{r}{4}$, and $u_l \stackrel{\Delta}{=} (C_1 + 1)$ for $1 \leq l \leq k$, to find that there exists a constant $K_r < \infty$ (depending only on r) s.t. for all $x > 0$,

$$\mathbb{P}\big(\max_{i=1,\ldots,k}(W_{n,i} - ia_n) > x\big) \leq K_r(C_1+1)^{\frac{r}{2}} k^{\frac{r}{2}} x^{-r}. \tag{12}$$

It follows that assumption (ii) of Theorem 5 holds as well, with (in the notation of Theorem 5) $C_2 \stackrel{\Delta}{=} K_r(C_1+1)^{\frac{r}{2}}$, $\epsilon \stackrel{\Delta}{=} r - 2$. Combining the above, we find that all assumptions of Theorem 5 hold, and thus we may apply Theorem 5 to find that

$$\left\{ \frac{B}{n^{\frac{1}{2}} - B} \sup_{k \geq 0} \big(k - \sum_{i=1}^{n} N_i(\lambda_n^{-1} \sum_{j=1}^{k} A_j^1)\big), n \geq 1 \right\}$$

is tight. Combining with (4) completes the proof of Theorem 1. □

6. Large deviation results and proof of Theorem 2. In this section, we complete the proofs of our main results. We proceed by combining our upper and lower bounds with several known weak convergence results for (pooled) renewal processes and the suprema of Gaussian processes. 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 $\big(Z(t_1), \ldots, Z(t_k)\big)$ 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 $\mathbb{E}[Z(s)Z(t)]$, and refer the reader to [13],[19],[2],[29], and the references therein for details on existence, continuity, etc.

6.1. *Preliminary weak convergence results.* In this subsection we review several weak convergence results for renewal processes, and apply them to $A_n(t)$ and $\sum_{i=1}^{n} N_i(t)$. 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 [49]. Let $\mathcal{A}(t)$ denote the w.p.1 continuous Gaussian process s.t. $\mathbb{E}[\mathcal{A}(t)] = 0, \mathbb{E}[\mathcal{A}(s)\mathcal{A}(t)] = \mu c_A^2 \min(s,t)$, namely $\mathcal{A}(t)$ is a driftless Brownian motion. Then it follows from the well-known Functional Central Limit Theorem (FCLT) for renewal processes (see [5] Theorem 14.6) that

THEOREM 6. *For any $T \in [0, \infty)$, the sequence of processes $\{\lambda_n^{-\frac{1}{2}}\big(A_n(t) - \lambda_n \mu t\big)_{0 \leq t \leq T}, n \geq 1\}$ converges weakly to $\mathcal{A}(t)_{0 \leq t \leq T}$ in the space $D[0,T]$ under the J_1 topology.*

We now give a weak convergence result for $\sum_{i=1}^n N_i(t)$, which is stated in [49] (see Theorem 7.2.3) and formally proven in [47] (see Theorem 2).

THEOREM 7. *There exists a w.p.1 continuous Gaussian process $\mathcal{D}(t)$ s.t. $\mathbb{E}[\mathcal{D}(t)] = 0, \mathbb{E}[\mathcal{D}(s)\mathcal{D}(t)] = \mathbb{E}[\big(N_1(s) - \mu s\big)\big(N_1(t) - \mu t\big)]$ for all $s, t \geq 0$. Furthermore, for any $T \in [0, \infty)$, the sequence of processes $\{n^{-\frac{1}{2}}\big(\sum_{i=1}^n N_i(t) - n\mu t\big)_{0\leq t\leq T}, n \geq 1\}$ converges weakly to $\mathcal{D}(t)_{0\leq t\leq T}$ in the space $D[0,T]$ under the J_1 topology.*

We note that the T_0 assumptions (i) and (iii), which guarantee that $\mathbb{E}[S^{2+\epsilon}] < \infty$ and $\limsup_{x\downarrow 0} x^{-1}\mathbb{P}(S \leq x) < \infty$, ensure that the technical conditions required to apply [49] Theorem 7.2.3, namely that $E[S^2] < \infty$ and $\limsup_{x\downarrow 0} x^{-1}\big(\mathbb{P}(S \leq x) - \mathbb{P}(S = 0)\big) < \infty$, hold.

It follows from Theorems 6 - 7 that

LEMMA 6. *For any fixed $T \geq 0$, $\{n^{-\frac{1}{2}}\big(A_n(t) - \sum_{i=1}^n N_i(t)\big)_{0\leq t\leq T}, n \geq 1\}$ converges weakly to $\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big)_{0\leq t\leq T}$ in the space $D[0,T]$ under the J_1 topology.*

PROOF. Note that

$$n^{-\frac{1}{2}}\big(A_n(t) - \sum_{i=1}^n N_i(t)\big)_{0\leq t\leq T} = \left(\lambda_n^{\frac{1}{2}} n^{-\frac{1}{2}}\big(A_n(t) - \lambda_n\mu t\big)\lambda_n^{-\frac{1}{2}} - \big(\sum_{i=1}^n N_i(t) - n\mu t\big)n^{-\frac{1}{2}} - B\mu t\right)_{0\leq t\leq T}.$$

The lemma then follows from Theorems 6 - 7. □

We note that a process very similar to $\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big)_{0\leq t\leq T}$ was studied in [47] as the weak limit of a sequence of queues with superposition arrival processes. The continuity of the supremum map in the space $D[0,T]$ under the J_1 topology (see [49] Theorem 13.4.1), combined with Lemma 6, implies that

COROLLARY 2. *For any fixed $T \geq 0$, $\{n^{-\frac{1}{2}}\sup_{0\leq t\leq T}\big(A_n(t) - \sum_{i=1}^n N_i(t)\big), n \geq 1\}$ converges weakly to the r.v. $\sup_{0\leq t\leq T}\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big)$.*

6.2. *Preliminary large deviation results.* Before proceeding with the remaining proofs, we will need to establish some results from the theory of large deviations of Gaussian processes and their suprema. We note that the relationship between the large deviations of suprema of Gaussian processes and the large deviations of queueing systems is well known (see [14], [12]), and there is a significant literature studying the large deviations of such processes (e.g. [14], [25], [24], [11], [12]). We will rely heavily on the following theorem, proven in [12] (in a more general form).

THEOREM 8. *Suppose $\mathcal{Z}(t)$ is a centered, separable Gaussian process with stationary increments, s.t. $E[\mathcal{Z}^2(t)]$ is a continuous function of t on $[0,\infty)$, $\lim_{t\downarrow 0}\big(E[\mathcal{Z}^2(t)]\log^2(t)\big) = 0$, and $\lim_{t\to\infty} t^{-1}\mathbb{E}[\mathcal{Z}^2(t)] = \sigma^2 > 0$. Then for any $c > 0$,*

$$\lim_{x\to\infty} x^{-1}\log\left(\mathbb{P}\big(\sup_{t\geq 0}(\mathcal{Z}(t) - ct) \geq x\big)\right) = -\frac{2c}{\sigma^2}.$$

It is also implicit from the discussion in [14] (although we include a short proof) that

THEOREM 9. *Under the same assumptions as Theorem 8, for any $c > 0$,*

$$\lim_{x\to\infty} x^{-1} \log\left(\sup_{t\geq 0} \mathbb{P}\big(\mathcal{Z}(t) - ct > x\big)\right) = -\frac{2c}{\sigma^2}.$$

PROOF. That $\limsup_{x\to\infty} x^{-1} \log\left(\sup_{t\geq 0} \mathbb{P}\big(\mathcal{Z}(t) - ct > x\big)\right) \leq -\frac{2c}{\sigma^2}$ follows immediately from Theorem 8 and the fact that $\sup_{t\geq 0} \mathbb{P}\big(\mathcal{Z}(t) - ct > x\big) \leq \mathbb{P}\Big(\sup_{t\geq 0} \big(\mathcal{Z}(t) - ct\big) > x\Big)$.

Letting $t = \frac{x}{c}$, we find that

$$\sup_{t\geq 0} \mathbb{P}\big(\mathcal{Z}(t) - ct > x\big) \geq \mathbb{P}\big(\mathcal{Z}(\frac{x}{c}) - x > x\big). \tag{13}$$

Let G denote a normally distributed r.v. with mean 0 and variance 1. Then since $\mathcal{Z}(\frac{x}{c})$ is normally distributed with mean zero, it follows from (13) that

$$\sup_{t\geq 0} \mathbb{P}\big(\mathcal{Z}(t) - ct > x\big) \geq \mathbb{P}\Big(G > 2x\mathbb{E}^{-\frac{1}{2}}\big[\mathcal{Z}^2(\frac{x}{c})\big]\Big). \tag{14}$$

We use the following identity from [1] Equation 7.1.13. Namely, for all $y > 0$,

$$\mathbb{P}(G > y) \geq \big(y + (y^2+4)^{-\frac{1}{2}}\big)^{-1} (\frac{2}{\pi})^{\frac{1}{2}} \exp(-\frac{y^2}{2}).$$

Thus

$$\mathbb{P}(G > y) \geq \exp(-\frac{y^2}{2} - y) \text{ for all sufficiently large } y. \tag{15}$$

By assumption, $\lim_{t\to\infty} t^{-1}\mathbb{E}[\mathcal{Z}^2(t)] = \sigma^2 > 0$, and thus $\lim_{x\to\infty} 2x\mathbb{E}^{-\frac{1}{2}}\big[\mathcal{Z}^2(\frac{x}{c})\big] = \infty$. It thus follows from (14) and (15) that for all sufficiently large x,

$$x^{-1} \log\left(\sup_{t\geq 0} \mathbb{P}\big(\mathcal{Z}(t) - ct > x\big)\right) \geq -2x\mathbb{E}^{-1}\big[\mathcal{Z}^2(\frac{x}{c})\big] - 2\mathbb{E}^{-\frac{1}{2}}\big[\mathcal{Z}^2(\frac{x}{c})\big].$$

Since $\lim_{x\to\infty}(\frac{x}{c})^{-1}\mathbb{E}[\mathcal{Z}^2(\frac{x}{c})] = \sigma^2$, it follows that $\liminf_{x\to\infty} x^{-1} \log\left(\sup_{t\geq 0} \mathbb{P}\big(\mathcal{Z}(t) - ct > x\big)\right) \geq -\frac{2c}{\sigma^2}$, concluding the proof of the theorem. □

In light of Theorem 8, Theorem 9 can be interpreted as saying that such a process is 'most likely' to exceed a given value x at a particular time (roughly $\frac{x}{c}$), and much less likely to exceed that value at any other time (see the discussion in [14]). We note that the duality of Theorems 8 - 9 coincides with the duality exhibited by our upper and lower bounds (Theorems 3 - 4) - a relationship that we will exploit to prove our large deviation results.

We are now in a position to apply Theorems 8 - 9 to $\mathcal{A}(t) - \mathcal{D}(t)$.

COROLLARY 3.

(i) $\lim_{x\to\infty} x^{-1} \log \mathbb{P}\Big(\sup_{t\geq 0} \big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big) \geq x\Big) = -2B(c_A^2 + c_S^2)^{-1}$;

(ii) $\lim_{x\to\infty} x^{-1}\log\Big(\sup_{t\geq 0}\mathbb{P}\big(\mathcal{A}(t)-\mathcal{D}(t)-B\mu t>x\big)\Big)=-2B(c_A^2+c_S^2)^{-1}$.

PROOF. That $\mathcal{A}(t)-\mathcal{D}(t)$ is a centered, separable Gaussian process with stationary increments follows from definitions, the independence of $\mathcal{A}(t)$ and $\mathcal{D}(t)$, and the fact that both $\mathcal{A}(t)$ and $N_1(t)$ have stationary increments. The independence of $\mathcal{A}(t)$ and $\mathcal{D}(t)$ implies that

$$\mathbb{E}[\big(\mathcal{A}(t)-\mathcal{D}(t)\big)^2]=\mu c_A^2 t+\mathbb{E}[\big(N_1(t)-\mu t\big)^2]. \tag{16}$$

Note that for all $t\geq 0$ and $k\geq 1$, $\mathcal{P}(N_1(t)\geq k)\leq \mathcal{P}\big(R(S)\leq t\big)\big(\mathcal{P}(S\leq t)\big)^{k-1}$. It thus follows from the T_0 assumptions (and a straightforward analogy to an appropriate geometrically distributed r.v.) that $\limsup_{t\downarrow 0} t^{-1}\mathbb{E}[\big(N_1(t)-\mu t\big)^2]<\infty$. Combining the above, we find that $\lim_{t\downarrow 0}\big(\mathbb{E}[\big(\mathcal{A}(t)-\mathcal{D}(t)\big)^2]\log^2(t)\big)=0$. In addition, the continuity of $\mathbb{E}[\big(\mathcal{A}(t)-\mathcal{D}(t)\big)^2]$ on $[0,\infty)$ follows from the above and a simple application of the Cauchy-Schwartz inequality.

Furthermore, we claim that $\lim_{t\to\infty} t^{-1}\mathbb{E}[\big(N_1(t)-\mu t\big)^2]=\mu c_S^2$. Indeed, let G_S denote a normally distributed r.v. with mean 0 and variance μc_S^2. It follows from the well-known Central Limit Theorem for renewal processes (see [39] Theorem 3.3.5), and the fact that $h(z)\stackrel{\Delta}{=}z^2$ is a continuous function, that the sequence of r.v.s $\{\Big(t^{-\frac{1}{2}}\big(N_1(t)-\mu t\big)\Big)^2, t\geq 1\}$ converges weakly to G_S^2. Recall that $\mathbb{E}[S^{2+\epsilon}]<\infty$ for some $\epsilon>0$ by the T_0 assumptions. Thus it follows from Lemma 5 that the sequence of r.v.s $\{\Big(t^{-\frac{1}{2}}\big(N_1(t)-\mu t\big)\Big)^2, t\geq 1\}$ is uniformly integrable. It follows that $\lim_{t\to\infty} t^{-1}\mathbb{E}[\big(N_1(t)-\mu t\big)^2]=\mu c_S^2$, since uniform integrability plus weak convergence implies convergence of moments.

Combining with (16), we find that $\lim_{t\to\infty} t^{-1}\mathbb{E}[\big(\mathcal{A}(t)-\mathcal{D}(t)\big)^2]=\mu(c_A^2+c_S^2)>0$ by the T_0 assumptions. It follows that $\mathcal{A}(t)-\mathcal{D}(t)$ satisfies the conditions needed to apply Theorems 8 - 9, from which the corollary follows. □

6.3. *Proof of Theorem 2.* Before completing the proofs of our main results, it will be useful to prove a strengthening of Theorem 1. Namely,

LEMMA 7. *For all $x\geq 0$,*

$$\lim_{T\to\infty}\limsup_{n\to\infty}\mathbb{P}\Big(n^{-\frac{1}{2}}\sup_{t\geq T}\big(A_n(t)-\sum_{i=1}^n N_i(t)\big)>x\Big)=0. \tag{17}$$

PROOF. Since $\mathbb{E}[A^{2+\epsilon}],\mathbb{E}[S^{2+\epsilon}]<\infty$ for some $\epsilon>0$ by the T_0 assumptions, we may fix some $r>2$ s.t. $\mathbb{E}[A^r],\mathbb{E}[S^r]<\infty$. Note that since $x\geq 0$, $\mathbb{P}\Big(n^{-\frac{1}{2}}\sup_{t\geq T}\big(A_n(t)-\sum_{i=1}^n N_i(t)\big)>x\Big)$ is at most $\mathbb{P}\Big(n^{-\frac{1}{2}}\sup_{t\geq T}\big(A_n(t)-\sum_{i=1}^n N_i(t)\big)>0\Big)$. By a simple union bound, $\mathbb{P}\Big(n^{-\frac{1}{2}}\sup_{t\geq T}\big(A_n(t)-\sum_{i=1}^n N_i(t)\big)>0\Big)$ is at most

$$\mathbb{P}\Big(n^{-\frac{1}{2}}\big(A_n(T)-\sum_{i=1}^n N_i(T)\big)>-\frac{B}{2}\mu T\Big) \tag{18}$$

$$+\mathbb{P}\Big(\sup_{t\geq T}\Big(n^{-\frac{1}{2}}\big(A_n(t)-\sum_{i=1}^n N_i(t)\big)-n^{-\frac{1}{2}}\big(A_n(T)-\sum_{i=1}^n N_i(T)\big)\Big)>\frac{B}{2}\mu T\Big). \tag{19}$$

We now bound (18), which equals

$$\begin{aligned}
& \mathbb{P}\Big(n^{-\frac{1}{2}} \big(A_n(T) - \lambda_n \mu T\big) - n^{-\frac{1}{2}} \big(\sum_{i=1}^{n} N_i(T) - n\mu T\big) - B\mu T > -\frac{B}{2}\mu T \Big) \\
\leq \; & \mathbb{P}\Big(|A_n(T) - \lambda_n \mu T| + |\sum_{i=1}^{n} N_i(T) - n\mu T| > n^{\frac{1}{2}} \frac{B}{2}\mu T \Big) \text{ by the tri. ineq.} \\
\leq \; & 2^{r-1} \Big(\mathbb{E}\big[|A_n(T) - \lambda_n \mu T|^r\big] + \mathbb{E}\big[|\sum_{i=1}^{n} N_i(T) - n\mu T|^r\big] \Big) n^{-\frac{r}{2}} \big(\frac{B}{2}\mu T\big)^{-r} \qquad (20) \\
& \text{by Markov's inequality and (6).}
\end{aligned}$$

Without loss of generality (w.l.o.g.) assuming $nT \geq \lambda_n T \geq 1$, it follows from Lemma 5 (applied with $n = 1$), and the fact that $A_n(T)$ has the same distribution as $A(\lambda_n T)$, that there exists $C_{A,r} \stackrel{\Delta}{=} \sup_{t \geq 1} t^{-\frac{r}{2}} \mathbb{E}\big[|A(t) - \mu t|^r\big] < \infty$ s.t.

$$\mathbb{E}\big[|A_n(T) - \lambda_n \mu T|^r\big] \leq C_{A,r} (\lambda_n T)^{\frac{r}{2}} \leq C_{A,r} (nT)^{\frac{r}{2}}. \qquad (21)$$

Since $nT \geq 1$ by assumption, it follows from Lemma 5 that there exist $C_{S,r} < \infty$ s.t.

$$\mathbb{E}\big[|\sum_{i=1}^{n} N_i(T) - n\mu T|^r\big] \leq C_{S,r} (nT)^{\frac{r}{2}}. \qquad (22)$$

It follows from (21) and (22) that (20) is at most

$$2^{r-1} (C_{A,r} + C_{S,r}) (\frac{B}{2}\mu)^{-r} T^{-\frac{r}{2}}.$$

Thus we find that

$$\lim_{T \to \infty} \limsup_{n \to \infty} \mathbb{P}\Big(n^{-\frac{1}{2}} \big(A_n(T) - \sum_{i=1}^{n} N_i(T)\big) > -\frac{B}{2}\mu T \Big) = 0. \qquad (23)$$

We now bound (19), which equals $\mathbb{P}\Big(n^{-\frac{1}{2}} \sup_{t \geq 0} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) > \frac{B}{2}\mu T \Big)$ by stationary increments. But as our proof of Theorem 1 demonstrates tightness of $\big\{ n^{-\frac{1}{2}} \sup_{t \geq 0} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big), n \geq 1 \big\}$, it follows that

$$\lim_{T \to \infty} \limsup_{n \to \infty} \mathbb{P}\Big(n^{-\frac{1}{2}} \sup_{t \geq 0} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) > \frac{B}{2}\mu T \Big) = 0. \qquad (24)$$

Using (24) to bound (19), we find that

$$\lim_{T \to \infty} \limsup_{n \to \infty} \mathbb{P}\Big(\sup_{t \geq T} \Big(n^{-\frac{1}{2}} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) - n^{-\frac{1}{2}} \big(A_n(T) - \sum_{i=1}^{n} N_i(T)\big) \Big) > \frac{B}{2}\mu T \Big) = 0. \qquad (25)$$

Combining (23) and (25) completes the proof. □

We now complete the proof of Theorem 2.

PROOF. We first prove the upper bound. By Lemma 7, for any $x > 0$, we may construct a strictly increasing sequence of integers $\{T_{x,k^{-1}}, k \geq 1\}$ s.t. for all $k \geq 1$,

$$\limsup_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{2}} \sup_{t \geq T_{x,k^{-1}}} \left(A_n(t) - \sum_{i=1}^{n} N_i(t) \right) \geq x \right) < k^{-1}.$$

It follows that for all $x > 0$ and $k \geq 1$,

(26)

$$\limsup_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{2}} \sup_{t \geq 0} \left(A_n(t) - \sum_{i=1}^{n} N_i(t) \right) \geq x \right) \leq \limsup_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{2}} \sup_{0 \leq t \leq T_{x,k^{-1}}} \left(A_n(t) - \sum_{i=1}^{n} N_i(t) \right) \geq x \right) + k^{-1}.$$

By the Portmanteau Theorem (see [5]), a sequence of r.v.s $\{X_n\}$ converges weakly to the r.v. X_∞ iff for all closed subsets C of $\mathbb{R}$, $\limsup_{n\to\infty} \mathbb{P}(X_n \in C) \leq \mathbb{P}(X_\infty \in C)$ iff for all open subsets O of $\mathbb{R}$, $\mathbb{P}(X_\infty \in O) \leq \liminf_{n\to\infty} \mathbb{P}(X_n \in O)$. It follows from (26) and Corollary 2 that for all $x > 0$ and $k \geq 1$,

$$\limsup_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{2}} \sup_{t \geq 0} \left(A_n(t) - \sum_{i=1}^{n} N_i(t) \right) \geq x \right) \leq \mathbb{P}\left(\sup_{0 \leq t \leq T_{x,k^{-1}}} \left(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t \right) \geq x \right) + k^{-1}. \quad (27)$$

Note that the sequence of events $\left\{ \sup_{0 \leq t \leq T_{x,k^{-1}}} \left(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t \right) \geq x, k \geq 1 \right\}$ is monotone in k. It follows that

$$\lim_{k\to\infty} \mathbb{P}\left(\sup_{0 \leq t \leq T_{x,k^{-1}}} \left(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t \right) \geq x \right) = \mathbb{P}\left(\sup_{t \geq 0} \left(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t \right) \geq x \right).$$

It then follows from (27), by letting $k \to \infty$, that for all $x > 0$,

$$\limsup_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{2}} \sup_{t \geq 0} \left(A_n(t) - \sum_{i=1}^{n} N_i(t) \right) \geq x \right) \leq \mathbb{P}\left(\sup_{t \geq 0} \left(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t \right) \geq x \right). \quad (28)$$

From Theorem 3 and (28) we have

$$\begin{aligned} & \limsup_{x\to\infty} x^{-1} \log \left(\limsup_{n\to\infty} \mathbb{P}\left(\left(Q^n(\infty) - n \right)^+ n^{-\frac{1}{2}} \geq x \right) \right) \\ \leq \; & \limsup_{x\to\infty} x^{-1} \log \mathbb{P}\left(\sup_{t \geq 0} \left(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t \right) \geq x \right) \\ = \; & -2B(c_A^2 + c_S^2)^{-1} \text{ by Corollary 3.(i),} \end{aligned}$$

which completes the proof of the upper bound.

We now complete the proof of Theorem 2 by demonstrating that if A is an exponentially distributed r.v., then

$$\liminf_{x\to\infty} x^{-1} \log \left(\liminf_{n\to\infty} \mathbb{P}\left(\left(Q^n(\infty) - n \right)^+ n^{-\frac{1}{2}} > x \right) \right) \geq -2B(c_A^2 + c_S^2)^{-1}. \quad (29)$$

Let Z_n denote a Poisson r.v. with mean λ_n. It follows from Theorem 4 that for all $x > 0$,

(30)
$$\liminf_{n\to\infty} \mathbb{P}\Big(\big(\mathrm{Q}^n(\infty) - n\big)^+ n^{-\frac{1}{2}} > x \Big) \geq \Big(\liminf_{n\to\infty} \mathbb{P}(Z_n \geq n) \Big) \Big(\liminf_{n\to\infty} \sup_{t\geq 0} \mathbb{P}\Big(n^{-\frac{1}{2}} \big(A_n(t) - \sum_{i=1}^{n} N_i(t) \big) > x \Big) \Big).$$

Recall that G is a normally distributed r.v. with mean 0 and variance 1. Thus by the Central Limit Theorem,

$$\lim_{n\to\infty} \mathbb{P}(Z_n \geq n) = \mathbb{P}\big(G \geq B\big). \tag{31}$$

Note that for any fixed t, $\mathcal{A}(t) - \mathcal{D}(t) - B\mu t$ is a non-degenerate Gaussian r.v., and every $x \in \mathbb{R}$ is a continuity point of the distribution of any non-degenerate Gaussian r.v. It follows from Lemma 6 and the definition of weak convergence that for any fixed $t \geq 0$ and all $x > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Big(n^{-\frac{1}{2}} \big(A_n(t) - \sum_{i=1}^{n} N_i(t) \big) > x \Big) = \mathbb{P}\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t > x \big).$$

Thus for any fixed $x > 0$ and $s \geq 0$,

$$\begin{aligned} \liminf_{n\to\infty} \sup_{t\geq 0} \mathbb{P}\Big(n^{-\frac{1}{2}} \big(A_n(t) - \sum_{i=1}^{n} N_i(t) \big) > x \Big) &\geq \liminf_{n\to\infty} \mathbb{P}\Big(n^{-\frac{1}{2}} \big(A_n(s) - \sum_{i=1}^{n} N_i(s) \big) > x \Big) \\ &= \mathbb{P}\big(\mathcal{A}(s) - \mathcal{D}(s) - B\mu s > x \big). \end{aligned} \tag{32}$$

By fixing $x > 0$ and taking the supremum over all $s \geq 0$ in (32), we find that for all $x > 0$,

$$\liminf_{n\to\infty} \sup_{t\geq 0} \mathbb{P}\Big(n^{-\frac{1}{2}} \big(A_n(t) - \sum_{i=1}^{n} N_i(t) \big) > x \Big) \geq \sup_{t\geq 0} \mathbb{P}\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t > x \big). \tag{33}$$

Combining (30), (31), and (33), we find that the left hand side (l.h.s.) of (30) is at least

$$\mathbb{P}\big(G \geq B\big) \sup_{t\geq 0} \mathbb{P}\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t > x \big). \tag{34}$$

(29) then follows from (34) and Corollary 3.ii. Combining (29) with the first part of Theorem 2, which we have already proven, completes the proof. □

7. Application to Reed's weak limit. In [37], J. Reed resolved the long-standing open question, originally posed in [18], of the tightness and weak convergence for the queue length of the transient $GI/GI/n$ queue in the H-W regime. However, the associated weak limit is only described implicitly, as the solution to a certain stochastic convolution equation (see [37]).

In this section we derive bounds for the weak limit of the transient $GI/GI/n$ queue in the H-W regime. Let $\mathcal{Q}_1^n$ denote the FCFS $GI/GI/n$ queue with inter-arrival times drawn i.i.d. distributed as $A\lambda_n^{-1}$, processing times drawn i.i.d. distributed as S, and the following initial conditions. For $i = 1, \ldots, n$, there is a single job initially being processed on server i, and the set of initial processing times of these n initial jobs is drawn i.i.d. distributed as $R(S)$; there are zero jobs waiting in queue, and the first inter-arrival time is distributed as $R(A\lambda_n^{-1})$, independent of the initial processing times of those jobs initially in system. Note that $\mathcal{Q}_1^n$ has the same initial conditions as the FCFS $GI/GI/n$ queue $\mathcal{Q}$ we considered when constructing our upper bound in Section 3. Let $\hat{Q}_1(t)$ denote the unique strong solution to the stochastic convolution equation given in [37] Equation 1.1. Then letting $Q_1^n(t)$ denote the number in system at time t in $\mathcal{Q}_1^n$, it is proven in [37] that

THEOREM 10. *For all $T \in (0,\infty)$, the sequence of stochastic processes $\{n^{-\frac{1}{2}}(Q_1^n(t)-n)^+_{0\le t\le T}, n \ge 1\}$ converges weakly to $\hat{Q}_1(t)_{0\le t\le T}$ in the space $D[0,T]$ under the J_1 topology.*

We now apply Theorem 3 to derive the first non-trivial bounds for $\hat{Q}_1(t)$, proving that

THEOREM 11. *For all $x > 0$ and $t \ge 0$,*

$$\mathbb{P}(\hat{Q}_1(t) > x) \le \mathbb{P}\bigg(\sup_{0\le s\le t}\big(\mathcal{A}(s) - \mathcal{D}(s) - B\mu s\big) \ge x\bigg).$$

PROOF. Note that we may let the arrival process to $\mathcal{Q}_1^n$ be $A_n(t)$. Thus by Theorem 3, for all $x > 0$ and $t \ge 0$,

$$\begin{aligned}
\liminf_{n\to\infty}\mathbb{P}\bigg(n^{-\frac{1}{2}}\big(\mathrm{Q}_1^n(t)-n\big)^+ > x\bigg) &\le \liminf_{n\to\infty}\mathbb{P}\bigg(n^{-\frac{1}{2}}\sup_{0\le s\le t}\big(A_n(s)-\sum_{i=1}^n N_i(s)\big) > x\bigg)\\
&\le \limsup_{n\to\infty}\mathbb{P}\bigg(n^{-\frac{1}{2}}\sup_{0\le s\le t}\big(A_n(s)-\sum_{i=1}^n N_i(s)\big) \ge x\bigg)\\
&\le \mathbb{P}\bigg(\sup_{0\le s\le t}\big(\mathcal{A}(s)-\mathcal{D}(s)-B\mu s\big) \ge x\bigg) \quad \text{by the Portmanteau Theorem.} \qquad (35)
\end{aligned}$$

Again applying the Portmanteau Theorem, it follows from Theorem 10 that for all $x > 0$,

$$\mathbb{P}(\hat{Q}_1(t) > x) \le \liminf_{n\to\infty}\mathbb{P}\bigg(n^{-\frac{1}{2}}\big(Q_1^n(t)-n\big)^+ > x\bigg). \qquad (36)$$

Combining (35) and (36) completes the proof. □

Theorem 11 implies that $\hat{Q}_1(t)$ is distributionally bounded over time, and thus in a sense stable. In particular, for all $t \ge 0$, $\hat{Q}_1(t)$ is stochastically dominated by the r.v. $\sup_{t\ge 0}\big(\mathcal{A}(t)-\mathcal{D}(t)-B\mu t\big)$.

8. Conclusion. In this paper, we studied the FCFS $GI/GI/n$ queue in the Halfin-Whitt regime. We proved that under minor technical conditions the associated sequence of steady-state queue length distributions, normalized by $n^{\frac{1}{2}}$, is tight. We derived an upper bound for the large deviation exponent of the limiting steady-state queue length matching that conjectured in [16], and proved a matching lower bound for the case of Poisson arrivals. We also derived the first non-trivial bounds for the weak limit process studied in [37]. Our main proof technique was the derivation of new and simple bounds for the FCFS $GI/GI/n$ queue, which are of a structural nature, and exemplify a general methodology which may be useful for analyzing a variety of queueing systems.

This work leaves many interesting directions for future research. One pressing question is whether or not $\{n^{-\frac{1}{2}}\big(Q^n(\infty)-n\big)^+, n \ge 1\}$ has a *unique* weak limit, and thus converges weakly. Indeed, such a result is only known for the cases of Markovian processing times [18], deterministic processing times [23], and processing times with finite support [16]. In all of these cases, either the distribution of $Q^n(\infty)$ can be computed explicitly [18],[23], or can be represented as the steady-state of a Markov chain whose dimension does not grow with n [16]; in the general setting, neither of these conditions hold. Similarly, although Theorem 11 shows that the weak limit process $\hat{Q}_1(t)$ is distributionally bounded over time, it is unknown whether $\hat{Q}_1(t)$ has a well-defined stationary measure. Furthermore, should $\{n^{-\frac{1}{2}}\big(Q^n(\infty)-n\big)^+, n \ge 1\}$ have a unique weak limit and $\hat{Q}_1(t)$ have

a well-defined stationary measure, must the two coincide? We note that this question is intimately related to showing that if one initializes $\mathcal{Q}^n$ with its stationary measure, then the relevant sequence of scaled queueing processes converges weakly (at the process level) to an appropriate stationary limit, and refer the reader to [42], [18], [33], [21], [35] for progress along these lines. Similar questions (on the order of fluid, as opposed to diffusion, scaling) were also investigated in [26].

It would be interesting to extend our techniques to more general models. For example, it should be possible to extend our lower bounds to non-Poisson arrival processes, as was done in [16] for the special case of processing times with finite support. It would also be interesting to analyze the large deviation behavior when the finite second moment condition does not hold, since in this case the large deviation exponent of Theorem 2 equals zero, which suggests that a fundamentally different qualitative behavior may arise in this setting. Finally, it would be interesting to generalize our bounds to systems with abandonments ($GI/GI/n + GI$). This setting is practically important, as the main application of the H-W regime has been to the study of call-centers, for which customer abandonments are an important modeling component [3]. For some interesting steps along these lines the reader is referred to the recent papers [10],[17].

9. Appendix.

9.1. *Proof of Lemma 2.*

PROOF. Let $Z^i(t)$ denote the number of jobs initially in $\mathcal{Q}^i$ which are still in $\mathcal{Q}^i$ at time t, $i \in \{1, 2\}$. We claim that $Z^2(t) \geq Z^1(t)$ for all $t \geq 0$. Indeed, let J be any job initially in $\mathcal{Q}^1$, and let S_J denote its initial processing time. Then (ii) ensures the existence of a distinct corresponding job J' initially in $\mathcal{Q}^2$, with the same initial processing time S_J. Since by (i) all jobs initially in Q^1 begin processing at time 0, it follows that J departs $\mathcal{Q}^1$ at time S_J, while J' departs $\mathcal{Q}^2$ no earlier than S_J. Making this argument for each job J initially in $\mathcal{Q}^1$ proves that $Z^2(t) \geq Z^1(t)$ for all $t \geq 0$.

Let D_k^i denote the time at which the job that arrives to $\mathcal{Q}^i$ at time T_k^1 departs from $\mathcal{Q}^i$, $k \geq 1, i \in \{1, 2\}$. We now prove by induction that for $k \geq 1$, $D_k^2 \geq D_k^1$, from which the proposition follows. Observe that for all $k \geq 1$,

$$D_k^1 = \inf\{t : \ t \geq T_k^1, Z^1(t) + \sum_{j=1}^{k-1} I(D_j^1 > t) \leq n - 1\} + S_k^1. \tag{37}$$

Also,

$$D_k^2 \geq \inf\{t : \ t \geq T_k^1, Z^1(t) + \sum_{j=1}^{k-1} I(D_j^2 > t) \leq n - 1\} + S_k^1, \tag{38}$$

where the inequality in (38) arises since $Z^2(t) \geq Z^1(t)$ for all $t \geq 0$, and the job that arrives to $\mathcal{Q}^2$ at time T_k^1 may have to wait for additional jobs, which either were initially present in $\mathcal{Q}^2$ but not $\mathcal{Q}^1$, or which arrive at a time belonging to $\{T_k^2, k \geq 1\} \setminus \{T_k^1, k \geq 1\}$.

For the base case $k = 1$, note that $D_1^1 = \inf\{t : t \geq T_1^1, Z^1(t) \leq n - 1\} + S_1^1$, while $D_1^2 \geq \inf\{t : t \geq T_1^1, Z^1(t) \leq n - 1\} + S_1^1$.

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

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

$$\leq \inf\{t:\ t \geq T^1_{k+1}, Z^1(t) + \sum_{j=1}^{k} I(D^2_j > t) \leq n-1\} + S^1_{k+1}.$$

It then follows from (37) and (38) that $D^1_{k+1} \leq D^2_{k+1}$, completing the induction. □

9.2. *Proof of Theorem 5.* It is proven in [44] Theorem 1 (given in the notation of [44]) that

THEOREM 12. *Suppose that for all sufficiently large n, $\{\zeta_{n,i}, i \geq 1\}$ is a stationary, countably infinite sequence of r.v. Let $a_n \stackrel{\Delta}{=} \mathbb{E}\big[\zeta_{n,1}\big]$, and $W_{n,k} \stackrel{\Delta}{=} \sum_{i=1}^{k} \zeta_{n,i}$. Further assume that $a_n < 0, \lim_{n\to\infty} a_n = 0$, and there exist $C_1, C_2 < \infty$ and $\epsilon > 0$ s.t. for all sufficiently large n,*

(i) $\mathbb{E}\big[|W_{n,k} - ka_n|^{2+\epsilon}\big] \leq C_1 k^{1+\frac{\epsilon}{2}}$ *for all* $k \geq 1$;
(ii) $\mathbb{P}\big(\max_{i=1,\ldots,k}(W_{n,i} - ia_n) > x\big) \leq C_2 \mathbb{E}\big[|W_{n,k} - ka_n|^{2+\epsilon}\big] x^{-(2+\epsilon)}$ *for all* $k \geq 1$ *and* $x > 0$;
(iii) $\mathbb{P}(\lim_{k\to\infty} W_{n,k} = -\infty) = 1$.

Then $\{|a_n| \sup_{k\geq 0} W_{n,k}, n \geq 1\}$ *is tight.*

With Theorem 12 in hand, we now complete the proof of Theorem 5.

PROOF OF THEOREM 5. The proof follows almost exactly as the proof of Theorem 12 given in [44], and we now explicitly comment on precisely where the proof must be changed superficially so as to carry through under the slightly different set of assumptions of Theorem 5. First off, nowhere in the proof of Theorem 12 given in [44] is assumption (iii) of Theorem 12 used, and thus that assumption is extraneous and may be removed. The only other difference between the set of assumptions for Theorem 12 and the set of assumptions for Theorem 5 is that assumption (ii) of Theorem 12 is replaced by assumption (ii) of Theorem 5. We now show that Theorem 12 holds under this change in assumptions. As in [44], let $x(a_n, k) \stackrel{\Delta}{=} \frac{x}{|a_n|} + 2^k|a_n|$. Then the only place where assumption (ii) of Theorem 12 is used is between Equations 5 and 6, where this assumption is required to demonstrate that

$$\mathbb{P}\big(W_{n,2^k} - 2^k a_n > \frac{1}{2}x(a_n,k)\big) + \mathbb{P}\big(\max_{i=0,\ldots,2^k}(\sum_{j=1}^{i} \zeta_{n,j+2^k} - ia_n) > \frac{1}{2}x(a_n,k)\big) \tag{39}$$

$$\leq (1+C_2)C_1 2^{2+\epsilon} 2^{k(1+\frac{\epsilon}{2})} \big(x(a_n,k)\big)^{-(2+\epsilon)}. \tag{40}$$

We now prove that assumption (ii) of Theorem 5 is sufficient to derive (40). In particular, the first summand of (39) is at most

$$\mathbb{E}\big[|W_{n,2^k} - 2^k a_n|^{2+\epsilon}\big] \big(\frac{1}{2}x(a_n,k)\big)^{-(2+\epsilon)} \quad \text{by Markov's inequality}$$

$$\leq C_1 2^{2+\epsilon} 2^{k(1+\frac{\epsilon}{2})} \big(x(a_n,k)\big)^{-(2+\epsilon)} \quad \text{by assumption (i) of Theorem 5.} \tag{41}$$

By the stationarity of $\{\zeta_{n,i}, i \geq 1\}$, the second summand of (39) equals

$$\mathbb{P}\big(\max_{i=0,\ldots,2^k}(W_{n,i} - ia_n) > \frac{1}{2}x(a_n,k)\big)$$

$$\leq C_2 2^{2+\epsilon} 2^{k(1+\frac{\epsilon}{2})} \big(x(a_n,k)\big)^{-(2+\epsilon)} \quad \text{by assumption (ii) of Theorem 5.} \tag{42}$$

Since we may w.l.o.g. take $C_1, C_2 \geq 1$, it follows that $C_1 + C_2 \leq (1+C_2)C_1$, and thus (40) follows from (41) and (42). The theorem follows from the proof of Theorem 12 given in [44]. □

9.3. *Proof of Lemma 5.* We note that the special case $r = 2$ is treated in [47]. Before proceeding with the proof of Lemma 5, it will be useful to prove three auxiliary results. The first treats the special case $n = 1, t \geq 1$ for ordinary (as opposed to equilibrium) renewal processes, and is proven in Theorem 1 of [8].

THEOREM 13. *Suppose $Z(t)$ is an ordinary renewal process with renewal distribution X s.t. $\mathbb{E}[X] = \mu^{-1} \in (0, \infty)$, and $\mathbb{E}[X^r] < \infty$ for some $r \geq 2$. Then $\sup_{t \geq 1} t^{-\frac{r}{2}} \mathbb{E}\big[|Z(t) - \mu t|^r\big] < \infty$.*

Second, we prove a lemma treating the special case $n = 1, t \geq 1$ for equilibrium renewal processes.

LEMMA 8. *Under the same definitions and assumptions as Lemma 5, for each $r \geq 2$, there exists $C_{X,r} < \infty$ (depending only on X and r) s.t. for all $t \geq 1$, $\mathbb{E}\big[|Z_1^e(t) - \mu t|^r\big] < C_{X,r} t^{\frac{r}{2}}$.*

PROOF. Let X^e denote the first renewal interval in $Z_1^e(t)$, and f_{X^e} its density function, whose existence is guaranteed by (1). Observe that we may construct $Z_1^e(t)$ and an ordinary renewal process $Z(t)$ (also with renewal distribution X) on the same probability space so that for all $t \geq 0$, $Z_1^e(t) = I(X^e \leq t) + Z\big((t - X^e)^+\big)$, with $Z(t)$ independent of X^e. Thus

$$Z_1^e(t) - \mu t = \Big(Z\big((t - X^e)^+\big) - \mu\big(t - X^e\big)^+\Big) + \Big(I\big(X^e \leq t\big) - \mu\big(t - (t - X^e)^+\big)\Big).$$

Fixing some $t \geq 1$, it follows that $\mathbb{E}\big[|Z_1^e(t) - \mu t|^r\big]$ is at most

$$2^{r-1}\mathbb{E}\big[|Z\big((t - X^e)^+\big) - \mu\big(t - X^e\big)^+|^r\big] \tag{43}$$

$$+2^{r-1}\mathbb{E}\big[|I\big(X^e \leq t\big) - \mu\big(t - (t - X^e)^+\big)|^r\big] \text{ by the triangle inequality and (6).} \tag{44}$$

We now bound the term $\mathbb{E}\big[|Z\big((t - X^e)^+\big) - \mu\big(t - X^e\big)^+|^r\big]$ appearing in (43), which equals

$$\int_0^{t-1} \mathbb{E}\big[|Z\big(t - s\big) - \mu\big(t - s\big)|^r\big] f_{X^e}(s) ds + \int_{t-1}^{t} \mathbb{E}\big[|Z\big(t - s\big) - \mu\big(t - s\big)|^r\big] f_{X^e}(s) ds. \tag{45}$$

Let $C'_{X,r} \stackrel{\Delta}{=} \sup_{t \geq 1} t^{-\frac{r}{2}} \mathbb{E}\big[|Z(t) - \mu t|^r\big]$. Theorem 13 implies that the first summand of (45) is at most

$$\begin{aligned} \int_0^{t-1} \big(C'_{X,r}(t - s)^{\frac{r}{2}}\big) f_{X^e}(s) ds &\leq \int_0^{t-1} \big(C'_{X,r} t^{\frac{r}{2}}\big) f_{X^e}(s) ds \\ &= C'_{X,r} t^{\frac{r}{2}} \mathbb{P}\big(X^e \leq t - 1\big). \end{aligned}$$

Since $t - s \leq 1$ implies $|Z(t - s) - \mu(t - s)|^r \leq |Z(1) + \mu|^r$, the second summand of (45) is at most $\mathbb{E}\big[|Z(1) + \mu|^r\big] \mathbb{P}\big(X^e \in [t - 1, t]\big)$. Combining our bounds for (45), we find that (43) is at most

$$2^{r-1}\mathbb{E}\big[|Z(1) + \mu|^r\big] + 2^{r-1} C'_{X,r} t^{\frac{r}{2}}. \tag{46}$$

We now bound (44), which is at most

$$\begin{aligned} 2^{r-1}\mathbb{E}\big[|I\big(X^e \leq t\big) + \mu\big(t - (t - X^e)^+\big)|^r\big] &\leq 2^{2r-2}\Big(1 + \mathbb{E}\big[|\mu\big(t - (t - X^e)^+\big)|^r\big]\Big) \text{ by (6)} \\ &= 2^{2r-2}\Big(1 + \mu^r\big(\int_0^t s^r f_{X^e}(s) ds + \int_t^\infty t^r f_{X^e}(s) ds\big)\Big). \end{aligned} \tag{47}$$

It follows from (1) and Markov's inequality that for all $s \geq 0$, $f_{X^e}(s) = \mu \mathbb{P}(X > s) \leq \mu \mathbb{E}[X^r] s^{-r}$. Thus the term $\int_0^t s^r f_{X^e}(s) ds + \int_t^\infty t^r f_{X^e}(s) ds$ appearing in (47) is at most

$$\begin{aligned} \int_0^t s^r \big(\mu \mathbb{E}[X^r] s^{-r}\big) ds + t^r \int_t^\infty \big(\mu \mathbb{E}[X^r] s^{-r}\big) ds &= \mu \mathbb{E}[X^r] \bigg(\int_0^t ds + t^r \int_t^\infty s^{-r} ds \bigg) \\ &= \mu \mathbb{E}[X^r] \bigg(t + t^r (r-1)^{-1} t^{1-r} \bigg) \\ &= \mu \mathbb{E}[X^r] \big(1 + (r-1)^{-1}\big) t. \end{aligned} \tag{48}$$

Using (46) to bound (43) and (48) to bound (47) and (44), we find that $\mathbb{E}\big[|Z_1^e(t) - \mu t|^r\big]$ is at most

$$2^{r-1} \mathbb{E}\big[|Z(1) + \mu|^r\big] + 2^{r-1} C'_{X,r} t^{\frac{r}{2}} + 2^{2r-2} + 2^{2r-2} \mu^{r+1} \mathbb{E}[X^r] \big(1 + (r-1)^{-1}\big) t. \tag{49}$$

Noting that $\mathbb{E}\big[|Z(1) + \mu|^r\big] < \infty$ since any renewal process, evaluated at any fixed time, has finite moments of all orders (see [30] p. 155), $\mathbb{E}[X^r] < \infty$ by assumption, and $t \leq t^{\frac{r}{2}}$ since $t \geq 1$ and $\frac{r}{2} \geq 1$, the lemma follows from (49). □

Third, we prove a lemma which will be useful in handling the case $t \leq 2$. We note that in this auxiliary lemma, the upper bound is of the form $(nt)^r$, as opposed to $(nt)^{\frac{r}{2}}$.

LEMMA 9. *Under the same definitions and assumptions as Lemma 5, there exists $C_{X,r} < \infty$ (depending only on X and r) s.t. for all $n \geq 1$, and $t \in \big[0, 2\big]$,*

$$\mathbb{E}\big[\big| \sum_{i=1}^n Z_i^e(t) - \mu n t\big|^r\big] \leq C_{X,r} \big(1 + (nt)^r\big). \tag{50}$$

PROOF. Note that the l.h.s. of (50) is at most

$$\mathbb{E}\big[|\sum_{i=1}^n Z_i^e(t) + \mu n t|^r\big] \leq 2^{r-1} \big(\mathbb{E}\big[\big(\sum_{i=1}^n Z_i^e(t)\big)^r\big] + (\mu n t)^r \big) \text{ by (6).} \tag{51}$$

We now bound the term $\mathbb{E}\big[\big(\sum_{i=1}^n Z_i^e(t)\big)^r\big]$ appearing in (51). Let $\{Z_i(t)\}$ denote a countably infinite sequence of i.i.d. ordinary renewal processes with renewal distribution X. Let us fix some $t \in [0, 2]$ and $n \geq 1$, and let $\{B_i\}$ denote a countably infinite sequence of i.i.d. Bernoulli r.v. s.t $\mathbb{P}(B_i = 1) = p \stackrel{\Delta}{=} \mathbb{P}(R(X) \leq t)$. Note that we may construct $\{Z_i^e(t)\}$, $\{Z_i(t)\}, \{B_i\}$ on the same probability space so that w.p.1 $Z_i^e(t) \leq B_i \big(1 + Z_i(t)\big)$ for all $i \geq 1$, with $\{Z_i(t)\}, \{B_i\}$ mutually independent. Letting $M \stackrel{\Delta}{=} \sum_{i=1}^n B_i$, it follows that

$$\mathbb{E}\big[\big(\sum_{i=1}^n Z_i^e(t)\big)^{\lceil r \rceil}\big] \leq \mathbb{E}\big[\bigg(\sum_{i=1}^M \big(1 + Z_i(t)\big)\bigg)^{\lceil r \rceil}\big]. \tag{52}$$

Let Z^+ denote the set of non-negative integers. Note that for any positive integer k,

$$\begin{aligned} \mathbb{E}\big[\bigg(\sum_{i=1}^k \big(1 + Z_i(t)\big)\bigg)^{\lceil r \rceil}\big] &= \mathbb{E}\big[\sum_{\substack{j_1, \ldots, j_k \in Z^+ \\ j_1 + \ldots + j_k = \lceil r \rceil}} \prod_{i=1}^k \big(1 + Z_i(t)\big)^{j_i}\big] \\ &= \sum_{\substack{j_1, \ldots, j_k \in Z^+ \\ j_1 + \ldots + j_k = \lceil r \rceil}} \prod_{i=1}^k \mathbb{E}\big[\big(1 + Z_i(t)\big)^{j_i}\big] \text{ since } \{Z_i(t)\} \text{ are i.i.d. r.v.s.} \end{aligned} \tag{53}$$

For any setting of $\{j_i, i = 1, \ldots, k\}$ in the r.h.s. of (53), at most $\lceil r \rceil$ of the j_i are strictly positive, and each j_i is at most $\lceil r \rceil$. It follows that the term $\prod_{i=1}^{k} \mathbb{E}\big[\big(1 + Z_i(t)\big)^{j_i}\big]$ appearing in the r.h.s. of (53) is at most $\left(\mathbb{E}\big[\big(1 + Z_1(t)\big)^{\lceil r \rceil}\big]\right)^{\lceil r \rceil}$, irregardless of the particular setting of $\{j_i, i = 1, \ldots, k\}$. As there are a total of $k^{\lceil r \rceil}$ distinct feasible configurations for $\{j_i, i = 1, \ldots, k\}$ in the r.h.s. of (53), combining the above we find that for any non-negative integer k,

$$\begin{aligned} \mathbb{E}\Big[\Big(\sum_{i=1}^{k} \big(1 + Z_i(t)\big)\Big)^{\lceil r \rceil}\Big] &\leq k^{\lceil r \rceil}\left(\mathbb{E}\big[\big(1 + Z_1(t)\big)^{\lceil r \rceil}\big]\right)^{\lceil r \rceil} \\ &\leq k^{\lceil r \rceil}\left(\mathbb{E}\big[\big(1 + Z_1(2)\big)^{\lceil r \rceil}\big]\right)^{\lceil r \rceil} \text{ since by assumption } t \leq 2. \end{aligned} \tag{54}$$

Since any renewal process, evaluated at any fixed time, has finite moments of all orders (see [30] p. 155), it follows that $C^1_{X,\lceil r \rceil} \stackrel{\Delta}{=} \left(\mathbb{E}\big[\big(1 + Z_1(2)\big)^{\lceil r \rceil}\big]\right)^{\lceil r \rceil}$ is a finite constant depending only on X and $\lceil r \rceil$. Combining (52) and (54) with the independence of M and $\{Z_i(t)\}$, it follows from a simple conditioning argument that

$$\mathbb{E}\big[\big(\sum_{i=1}^{n} Z^e_i(t)\big)^{\lceil r \rceil}\big] \leq C^1_{X,\lceil r \rceil} \mathbb{E}\big[M^{\lceil r \rceil}\big]. \tag{55}$$

We now bound the term $\mathbb{E}\big[M^{\lceil r \rceil}\big]$ appearing in (55). Noting that M is a binomial distribution with parameters n and p, it follows from [38] Equation 3.3 that there exist finite constants $C_{0,\lceil r \rceil}, C_{1,\lceil r \rceil}, C_{2,\lceil r \rceil}, \ldots, C_{\lceil r \rceil,\lceil r \rceil}$, independent of n and p, s.t. $\mathbb{E}\big[M^{\lceil r \rceil}\big] = \sum_{k=0}^{\lceil r \rceil} C_{k,\lceil r \rceil} p^k \prod_{j=0}^{k-1}(n-j)$. Further noting that $\prod_{j=0}^{k-1}(n-j) \leq n^k$ for all $k \geq 0$, it follows that $\mathbb{E}\big[M^{\lceil r \rceil}\big] \leq \sum_{k=0}^{\lceil r \rceil} |C_{k,\lceil r \rceil}|(np)^k$. Letting $C^2_{\lceil r \rceil} \stackrel{\Delta}{=} \max_{i=0,\ldots,\lceil r \rceil} |C_{i,\lceil r \rceil}|$, it follows from (55) that

$$\begin{aligned} \mathbb{E}\big[\big(\sum_{i=1}^{n} Z^e_i(t)\big)^{\lceil r \rceil}\big] &\leq C^1_{X,\lceil r \rceil} C^2_{\lceil r \rceil} \sum_{i=0}^{\lceil r \rceil} (np)^i \\ &\leq C^1_{X,\lceil r \rceil} C^2_{\lceil r \rceil} (\lceil r \rceil + 1)\big(1 + np\big)^{\lceil r \rceil}. \end{aligned} \tag{56}$$

Recall that for any non-negative r.v. Y, one has that $\mathbb{E}[Y^r] \leq \mathbb{E}[Y^{\lceil r \rceil}]^{\frac{r}{\lceil r \rceil}}$. Thus letting $C^3_{X,r} \stackrel{\Delta}{=} \left(C^1_{X,\lceil r \rceil} C^2_{\lceil r \rceil}(\lceil r \rceil + 1)\right)^{\frac{r}{\lceil r \rceil}}$, it follows from (56) that

$$\mathbb{E}\big[\big(\sum_{i=1}^{n} Z^e_i(t)\big)^{r}\big] \leq C^3_{X,r}(1 + np)^r. \tag{57}$$

Furthermore, it follows from (1) that $p = \mu \int_0^t \mathbb{P}(X > y)dy \leq \mu t$. Combining with (57), we find that

$$\mathbb{E}\big[\big(\sum_{i=1}^{n} Z^e_i(t)\big)^{r}\big] \leq C^3_{X,r}(1 + \mu n t)^r \tag{58}$$

Plugging (58) back into (51), it follows that the l.h.s. of (50) is at most

$$2^{r-1}\bigg(C_{X,r}^3(1+\mu nt)^r + (\mu nt)^r\bigg) \leq 2^r(C_{X,r}^3+1)(1+\mu nt)^r.$$

Noting that $(1+\mu nt)^r \leq 2^r\big(1+(\mu nt)^r\big)$ by (6), and $1+(\mu nt)^r \leq (1+\mu)^r\big(1+(nt)^r\big)$, completes the proof. □

With the above auxiliary results in hand, we now complete the proof of Lemma 5.

Proof of Lemma 5. We proceed by a case analysis. First, suppose $t \leq \frac{2}{n}$. Then we also have $t \leq 2$, and by Lemma 9 there exists $C_{X,r}^1 < \infty$ s.t. the l.h.s. of (7) is at most

$$C_{X,r}^1\big(1+(nt)^r\big) \quad \leq \quad C_{X,r}^1(1+2^r) \quad \text{ since } t \leq \frac{2}{n} \text{ implies } nt \leq 2.$$

Letting $M_1 \stackrel{\Delta}{=} C_{X,r}^1(1+2^r)$, it follows that the l.h.s. of (7) is at most $M_1 \leq M_1\big(1+(nt)^{\frac{r}{2}}\big)$, completing the proof for the case $t \leq \frac{2}{n}$.

Second, suppose $t \in \big[\frac{2}{n}, 2\big]$. Let $n_1(t) \stackrel{\Delta}{=} \lfloor nt \rfloor$. Noting that $t \geq \frac{2}{n}$ implies $n_1(t) > 0$, in this case we may define $n_2(t) \stackrel{\Delta}{=} \lfloor \frac{n}{n_1(t)} \rfloor$. Then the l.h.s. of (7) equals

$$\mathbb{E}\Big[\Big|\sum_{m=1}^{n_1(t)}\sum_{l=1}^{n_2(t)}\big(Z_{(m-1)n_2(t)+l}^e(t)-\mu t\big) + \sum_{l=n_1(t)n_2(t)+1}^{n}\big(Z_l^e(t)-\mu t\big)\Big|^r\Big]$$

$$\leq \quad 2^{r-1}\mathbb{E}\Big[\Big|\sum_{m=1}^{n_1(t)}\sum_{l=1}^{n_2(t)}\big(Z_{(m-1)n_2(t)+l}^e(t)-\mu t\big)\Big|^r\Big] \tag{59}$$

$$+2^{r-1}\mathbb{E}\Big[\Big|\sum_{l=n_1(t)n_2(t)+1}^{n}\big(Z_l^e(t)-\mu t\big)\Big|^r\Big] \quad \text{by the tri. ineq. and (6).} \tag{60}$$

We now bound (59). By Lemma 4, there exists $C_r < \infty$ s.t. (59) is at most

$$2^{r-1}C_r(n_1(t))^{\frac{r}{2}}\mathbb{E}\Big[\Big|\sum_{l=1}^{n_2(t)}\big(Z_l^e(t)-\mu t\big)\Big|^r\Big]$$

$$\leq \quad 2^{r-1}C_r(n_1(t))^{\frac{r}{2}}\bigg(C_{X,r}^1\Big(1+\big(n_2(t)t\big)^r\Big)\bigg) \text{ by Lemma 9, since } t \leq 2. \tag{61}$$

We now bound the term $tn_2(t)$ appearing in (61). In particular,

$$tn_2(t) \quad = \quad t\lfloor\frac{n}{\lfloor nt\rfloor}\rfloor \quad \leq \quad \frac{nt}{nt-1}. \tag{62}$$

But since $t \geq \frac{2}{n}$ imples $nt \geq 2$, and $g(z) \stackrel{\Delta}{=} \frac{z}{z-1}$ is a decreasing function of z on $(1,\infty)$, it follows from (62) that

$$tn_2(t) \leq 2.$$

Since $n_1(t) \le nt$, it thus follows from (61) that (59) is at most

$$2^{r-1}C_rC^1_{X,r}(1+2^r)(nt)^{\frac{r}{2}}. \tag{63}$$

We now bound (60). Note that the sum $\sum_{l=n_1(t)n_2(t)+1}^{n}\big(Z^e_l(t)-\mu t\big)$ appearing in (60) is taken over $n-n_1(t)n_2(t)$ terms. Furthermore,

$$\begin{aligned} n-n_1(t)n_2(t) &= n-n_1(t)\lfloor\frac{n}{n_1(t)}\rfloor \\ &\le n-n_1(t)\big(\frac{n}{n_1(t)}-1\big) \\ &= n_1(t). \end{aligned}$$

As $n_1(t) \le nt$, it thus follows from Lemma 4 that (60) is at most

$$\begin{aligned} 2^{r-1}C_r(nt)^{\frac{r}{2}}\mathbb{E}\big[|Z^e_1(t)-\mu t|^r\big] &\le 2^{r-1}C_r(nt)^{\frac{r}{2}}\mathbb{E}\big[\big(Z^e_1(t)+\mu t\big)^r\big] \\ &\le 2^{r-1}C_r(nt)^{\frac{r}{2}}\mathbb{E}\big[\big(Z^e_1(2)+2\mu\big)^r\big] \text{ since } t\le 2. \end{aligned} \tag{64}$$

Using (63) to bound (59) and (64) to bound (60) shows that the l.h.s. of (7) is at most

$$2^{r-1}C_rC^1_{X,r}(1+2^r)(nt)^{\frac{r}{2}} + 2^{r-1}C_r(nt)^{\frac{r}{2}}\mathbb{E}\big[\big(Z^e_1(2)+2\mu\big)^r\big]. \tag{65}$$

Let $M_2 \triangleq 2^{r-1}C_rC^1_{X,r}(1+2^r) + 2^{r-1}C_r\mathbb{E}\big[\big(Z^e_1(2)+2\mu\big)^r\big]$. It follows from (65) that the l.h.s. of (7) is at most $M_2(nt)^{\frac{r}{2}} \le M_2\big(1+(nt)^{\frac{r}{2}}\big)$, completing the proof for the case $t \in \big[\frac{2}{n},2\big]$.

Finally, suppose $t \ge 2$. In this case, it follows from Lemma 4 that the l.h.s. of (7) is at most $C_rn^{\frac{r}{2}}\mathbb{E}\big[|Z^e_1(t)-\mu t|^r\big]$. Let $C^2_{X,r} \triangleq \sup_{t\ge 2} t^{-\frac{r}{2}}\mathbb{E}\big[|Z^e_1(t)-\mu t|^r\big]$. Then it follows from Lemma 8 that $C^2_{X,r} < \infty$, and the l.h.s. of (7) is at most $C_rC^2_{X,r}(nt)^{\frac{r}{2}}$. Letting $M_3 \triangleq C_rC^2_{X,r}$, it follows that the l.h.s. of (7) is at most $M_3(nt)^{\frac{r}{2}} \le M_3\big(1+(nt)^{\frac{r}{2}}\big)$, completing the proof for the case $t \ge 2$.

As this treats all cases, we can complete the proof of the lemma by letting $M_4 \triangleq \max\big(M_1,M_2,M_3\big)$, and noting that for all $n \ge 1$ and $t \ge 0$, the l.h.s. of (7) is at most $M_4\big(1+(nt)^{\frac{r}{2}}\big)$. □

10. Acknowledgements. The authors would like to thank Ton Dieker, Johan van Leeuwaarden, Josh Reed, Ward Whitt, and Bert Zwart for their helpful discussions and insights. The authors especially thank Ton Dieker for his insights into the large deviations properties of suprema of Gaussian processes. The first author wishes to acknowledge support by NSF grant CMMI-0726733.

REFERENCES

[1] M. Abramowitz and I. Stegun, *Handbook of mathematical functions*, 1972.

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

[3] Z. Aksin, M. Armory, and V. Mehrotra, *The modern call center: a multi-disciplinary perspective on operations management research*, Production and Operations Management **16** (2007), no. 6, 665–688.

[4] S. Asmussen, *Applied probability and queues, 2nd ed.*, Springer, 2003.

[5] P. Billingsley, *Convergence of probability measures*, 1999.

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

[7] 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** (1994), 239–260.

[8] Y. Chao, C. Hsiung, and T. Lai, *Extended renewal theory and moment convergence in Anscombe's theorem*, The Annals of Probability **7** (1979), no. 2, 304–318.

[9] D. Cox, *Renewal theory*, Methuen and Co., 1970.

[10] J. Dai and S. He, *Customer abandonment in many-server queues.*, Mathematics of Operations Research. **35** (2010), 347–362.

[11] K. Debicki, *A note on LDP for supremum of Gaussian processes over infinite horizon*, Stat. Prob. Letters **44** (1999), 211–219.

[12] A.B. Dieker, *Conditional limit theorems for queues with Gaussian input, a weak convergence approach*, Stochastic processes and their applications **115** (2005), 849–873.

[13] J.L. Doob, *The elementary Gaussian processes*, Ann. Math. Statist. **15** (1944), no. 3, 229–282.

[14] N. Duffield and N. O'Connel, *Large deviations and overflow probabilities for the general single-server queue, with applications*, Math. Proc. Camb. Phil. Soc. **118** (1995).

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

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

[17] D. Gamarnik and A. Stolyar, *Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime. Asymptotics of the stationary distribution*, Preprint (2011).

[18] S. Halfin and W. Whitt, *Heavy-traffic limits for queues with many exponential servers.*, Operations Research **29** (1981), no. 3, 567–588.

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

[20] D. Iglehart and W. Whitt, *Multiple channel queues in heavy traffic I*, Advances in Applied Probability **2** (1970), no. 1, 150–177.

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

[22] D. Jagerman, *Some properties of the Erlang loss function*, Bell System Techn. J. **53** (1974), no. 3, 525–551.

[23] P. Jelenkovic, A. Mandelbaum, and P. Momcilovic, *Heavy traffic limits for queues with many deterministic servers*, Queueing Systems: Theory and Applications **47** (2004), no. 1-2, 53–69.

[24] Z. Michna K. Debicki and T. Rolski, *On the supremum from Gaussian processes over infinite horizon*, Probab. Math. Statist. **18** (1998), 83–100.

[25] J. Lewis K. Duffy and W. Sullivan, *Logarithmic asymptotics for the supremum of a stochastic process*, Annals of applied probability **13** (2003), 430–445.

[26] W. Kang and K. Ramanan, *Asymptotic approximations for the stationary distributions of many-server queues*, Annals of Applied Probability (to appear). (2010).

[27] O. Kella and W. Stadje, *Superposition of renewal processes and an application to multi-server queues*, Statistics and Probability Letters **76** (2006), no. 17, 1914–1924.

[28] A. Mandelbaum and P. Momcilovic, *Queues with many servers: the virtual waiting-time process in the QED regime*, Math. Oper. Res. **33** (2008), no. 3, 561–586.

[29] M. Marcus and J. Rosen, *Markov processes, Gaussian processes, and local times*, 2006.

[30] N. Prabhu, *Stochastic processes*, 1965.

[31] P. Protter, *Stochastic differential equations with jump reflection at the boundary*, Stochastics **3** (1980), 193–201.

[32] A. Puhalski and M. Reiman, *The multiclass GI/PH/N queue in the Halfin-Whitt regime*, Advances in Applied Probability **32** (2000), no. 3, 564–595.

[33] A.A. Puhalskii and M.I.Reiman, *The multiclass GI/PH/N queue in the halfin-whitt regime*, Advances in Applied Probability **32** (2000), 564 – 595.

[34] A.A. Puhalskii and J. Reed, *On many-server queues in heavy traffic.*, Ann. Appl. Prob. **20** (2010), no. 1, 129–195.

[35] K. Ramanan and H. Kaspi, *SPDE limits of many-server queues*, Preprint. (2010).

[36] J. Reed, *The G/GI/N queue in the Halfin-Whitt regime II: idle time system equations.*, Preprint. (2007).

[37] ———, *The G/GI/n queue in the Halfin-Whitt regime*, Annals of Applied Probability **19** (2009), no. 6, 2211–2269.

[38] J. Riordan, *Moment recurrence relations for binomial, Poisson and hypergeometric frequency distributions*, Ann. Math. Statist. **8** (1937), 103–111.

[39] S. Ross, *Stochastic processes, 2nd ed.*, Wiley and Sons, 1996.

[40] J. Shanthikumar and D. Yao, *Stochastic monotonicity in general queueing networks*, J. App. Prob. **26** (1989), 413–417.

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

[42] C. Stone, *Limit theorems for random walks, birth and death processes, and diffusion processes.*, Illinois J. Math. **4** (1963), 638–660.

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

[44] W. Szczotka, *Tightness of the stationary waiting time in heavy traffic*, Adv. Appl. Prob. **31** (1999), 788–794.

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

[46] W. Whitt, *Comparing counting processes and queues*, Advances in Applied Probability **13** (1981), no. 1, 207–220.

[47] ———, *Queues with superposition arrival processes in heavy traffic*, Stoch. Proc. App. **21** (1985), 81–91.

[48] ———, *The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting-time distribution*, Queueing Syst. Theory Appl. **36** (2000), no. 1/3, 71–87.

[49] ———, *Stochastic process limits*, Springer, 2002.

[50] P. Whittle, *Bounds for the moments of linear and quadratic forms in independent random variables*, Theor. Probability Appl. **5** (1960), 302–305.

Operations Research Center and Sloan School of Management, MIT, Cambridge, MA, 02139
E-mail: gamarnik@mit.edu
Georgia Institute of Technology, Atlanta, GA, 30332
E-mail: dgoldberg9@isye.gatech.edu