STOCHASTIC COMPARISON APPROACH TO MULTI-SERVER QUEUES: BOUNDS, HEAVY TAILS AND LARGE DEVIATIONS

A Dissertation
Presented to
The Academic Faculty

By

Yuan Li

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology

August 2017

STOCHASTIC COMPARISON APPROACH TO MULTI-SERVER QUEUES: BOUNDS, HEAVY TAILS AND LARGE DEVIATIONS

Approved by:

Professor David Goldberg, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Robert Foley
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Hayriye Ayhan
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Siva Theja Maguluri
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Jun Xu
School of Computer Science
Georgia Institute of Technology

Date Approved: May 09, 2017

This thesis is dedicated to my parents,

and to the memory of my grandpa.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitudes to my thesis advisor and mentor, Dr. David Goldberg, for his infinite patience, unwavering supports and guidance over my Ph.D. years, without whom this dissertation would not have been possible. I feel absolutely lucky working with him, an extraordinary researcher and a true inspiration.

Besides my advisor, I would like to thank the rest of my thesis committee members, Dr. Robert Foley, Dr. Hayriye Ayhan, Dr. Siva Theja Maguluri and Dr. Jun Xu, for their insightful comments, encouragements, and all the helps along the way.

Next I want to thank all my friends, for making my Ph.D. life colorful. I feel blessed to know them, Zihao Li, Zhihan Wang, Yuchen Zheng, Zhaocheng Li, Yilun Chen, Rui Gao, Can Zhang, Zhihao Ding, Amy Musselman, Richard Birge, Seunghan Lee, Zeinab Baharmi, Jikai Zou, Guido Lagos, Jan Vlachy... They are my families over the sea, my strengths to go through difficulties, and my biggest reward out of the Ph.D. program.

And my most heartfelt thanks to my parents. Words can not describe my gratitudes to them, for all their sacrifices, undying loves and supports. Mom and dad, thank you for making me fearless, giving me peace, and providing me so much freedom to do what I want. You are the reason I got this far.

TABLE OF CONTENTS

SUMMARY

The multi-server queue is one of the most studied subjects and celebrated tools in Operations Research, with manifold applications including manufacturing system, telecommunication networks, and homeland security. For a FCFS $GI/GI/n$ model queue, when the number of servers n is large, the direct analysis of various performance metrics is generally analytically intractable. Therefore methods like developing bounds and heavy-traffic approximation are crucial to help us understand the system. A particularly popular heavy-traffic regime is the so-called Halfin-Whitt regime, under which the number of servers n grows large and the traffic intensity ρ converges to unity simultaneously under some square-root staffing rule, allowing the system to strike a balance between quality and efficiency. However, all known bounds for general multi-server queues in this setting suffer from several fundamental shortcomings, such as holding only asymptotically or involving non-explicit constants. Furthermore, even less is known in the presence of heavy-tailed distributions in the model. This thesis primarily addresses these problems.

In the first part of this thesis, we consider the FCFS $GI/GI/n$ queue, and prove the first simple and explicit bounds that scale gracefully and universally as $\frac{1}{1-\rho}$ (ρ being the corresponding traffic intensity), across all notions of heavy traffic, including both the classical and Halfin-Whitt settings. In particular, supposing that the inter-arrival and service times, distributed as random variables A and S, have finite rth moment for some $r > 2$, and letting $\mu_A(\mu_S)$ denote $\frac{1}{\mathbb{E}[A]}(\frac{1}{\mathbb{E}[S]})$, our main results are bounds for the tail of the steady-state queue length and the steady-state probability of delay, expressed as simple and explicit functions of only $\mathbb{E}\big[(A\mu_A)^r\big], \mathbb{E}\big[(S\mu_S)^r\big], r$, and $\frac{1}{1-\rho}$.

In the second part of this thesis, we consider the FCFS $GI/GI/n$ queue in the Halfin-Whitt heavy traffic regime, in the presence of heavy-tailed distributions (i.e. infinite variance). We prove that under minimal assumptions, i.e. only that service times have finite $1+\epsilon$ moment for some $\epsilon > 0$ and inter-arrival times have finite second moment, the

sequence of stationary queue length distributions, normalized by $n^{\frac{1}{2}}$, is tight in the Halfin-Whitt regime. Furthermore, we develop simple and explicit bounds on the stationary queue length in that regime. For the setting where instead the inter-arrival times have an asymptotically Pareto tail with index $\alpha \in (1, 2)$, we extend recent results of [1] (who analyzed the case of deterministic service times) by proving that for general service time distributions, the sequence of stationary queue length distributions, normalized by $n^{\frac{1}{\alpha}}$, is tight (here we use the scaling of [1], which we refer to as the Halfin-Whitt-Reed scaling regime).

In the third part of this thesis, we further investigate the large deviation behaviors of the limits of the sequence of scaled stationary queue length in the presence of heavy-tailed inter-arrival and/or service times, which were proved to be tight previously. When service times have an asymptotically Pareto tail with index $\alpha \in (1, 2)$ and inter-arrival times have finite second moment, we bound the large deviation behavior of the ($n^{\frac{1}{2}}$-scaled) limiting process (defined as any suitable subsequential limit), and derive a matching lower bound when inter-arrival times are Markovian. Interestingly, we find that the large deviations behavior of the limit has a *sub-exponential* decay, differing fundamentally from the exponentially decaying tails known to hold in the light-tailed setting. For the setting where instead the inter-arrival times have an asymptotically Pareto tail with index $\alpha \in (1, 2)$, we are again able to bound the large-deviations behavior of the ($n^{\frac{1}{\alpha}}$-scaled and under Halfin-Whitt-Reed regime) limit, and find that our derived bounds do not depend on the particular service time distribution, and are in fact tight even for the case of deterministic service times.

CHAPTER 1
INTRODUCTION

1.1 Multi-server queues

The origins of queueing theory can be traced back to the work of T.O. Engset [2, 3] and A.K. Erlang [4, 5] on telecommunications in the early 1900s. While its definition can be as simple as "the study of the phenomena of standing, waiting, and serving" (Kleinrock [6]), queueing theory over the past hundred years has become one of the most studied areas and celebrated tools in Operations Research, with manifold applications including call centers [7, 8, 9]; homeland security [10, 11]; cloud computing [12, 13, 14]; and financial modeling [15, 16]. As a matter of fact, the applications of queueing theory are so rich, numerous surveys and bibliographies are published over the years, sometimes just for a specific field of application, e.g. manufacturing system[17, 18]; healthcare and medicine [19, 20].

As the bread-and-butter in the queueing theory, the multi-server queue has been a powerful modeling tool to capture complex dynamics of real-life problems, as well as the basic components for more advanced systems. But even as the most fundamental building block, the $GI/GI/n$ FCFS queue, in which jobs arrival independently following some random process, and wait in an un-capacitated waiting line until they are served (in their arriving order) by the next available server, few metrics are know without strong assumptions on the service time distribution, even for the $M/GI/n$ case. Tijms [21] commented that it is "not likely that computationally tractable methods can be developed to compute the exact numerical values of the steady-state probability in the $M/G/c$ queue ".

In general, as the number of servers n increases, the task to analyze the $GI/GI/n$ becomes increasingly difficult, maybe with an exception for the extreme case when $n = \infty$. The $GI/GI/\infty$ model exists not because there are infinite server queues observed in real

life, but rather due to the fact that it introduces insights (for large n), bounds, and the computational simplicity, e.g. mean queue length for $M/M/\infty$ queue has a very simple expression (being the traffic intensity) comparing to that of $M/M/n$ and the stationary distribution of the $M/GI/\infty$ behaves the same as that of $M/M/\infty$ [22]. Note that when $n = 2$, the problem (e.g. waiting time distribution) may also receive analytical remedy as it may be reduced to a functional equation which can be translated to some well-studied boundary value problems [23, 24].

However, in reality, the number of servers n may be large but finite, as demanded by real-life applications (e.g. when designing a call center [7, 8, 9]). Thus the exact analysis of performance metrics (e.g. mean stationary queue length and waiting time) are not in general analytically available, at least not without strong assumptions on service time distributions. Therefore, developing proper approximations and bounds are crucial to understanding the system, if not the only way to do so. There are three types of approximation following the classification in [25]. The first type of approximation utilizes bounds (e.g. [26]). By properly analyzing (and combining) upper and lower bounds, one may generate some useful ideas on the true solution of interest, and the (behavior of) gap between bounds hints the quality of the approximation. The second type is system approximation, and the idea is to approximate the original queueing system by a better-studied system. For example, this can be done by replacing the general service distribution in a $M/G/k$ queue with a phase-type distribution [27] or a discrete distribution [28] that mimics the original service time distribution as close as possible (generally through moments matching). The third type is process approximation, the idea of which is to properly scale (and center) the original process to generate a sequence of processes whose limit is another stochastic process that is generally well-studied (cf. [29]).

1.2 Bounds for multi-server queues

John Kingman in 1962 [30] proposed a simple and explicit upper bound (referred to as Kingman's bound) for the steady-state expected waiting time $\mathbb{E}[W]$ for a $GI/GI/1$ queue (which can be directly translated into the steady-state expected queue length (excluding jobs in service) $\mathbb{E}[Q]$), which states that

$$\mathbb{E}[W] \leq \frac{\sigma_A^2 + \sigma_S^2}{2\mathbb{E}[A]} \times \frac{1}{1-\rho} \quad , \quad \mathbb{E}[Q] \leq \frac{\sigma_A^2 + \sigma_S^2}{2(\mathbb{E}[A])^2} \times \frac{1}{1-\rho},$$

with $\sigma_A^2(\sigma_S^2)$ the variance of the inter-arrival (service time) distribution, $\mathbb{E}[A]$ the mean inter-arrival time, and ρ the traffic intensity. In the same paper Kingman also established that (under appropriate technical conditions) this bound becomes tight as $\rho \uparrow 1$.

As most performance metrics of the general $GI/GI/1$ queue have no simple closed-form solution, this combination of **simplicity, accuracy, and scalability** has made Kingman's bound very attractive over the years, from the perspective of both real queueing applications and as a theoretical tool. This motivates us to explore the question, does Kingman-type bound exist for multi-server queues?

To construct bounds for multi-server queues, one way is to establish error bounds on heavy-traffic approximations. The idea is to first establish heavy-traffic approximations of the true system and then bound the distance (error) between the approximation and the true system. Note that there are two popular ways $GI/GI/n$ queues can approach heavy traffic. The first is the classical heavy-traffic setting, in which the number of servers n is held fixed as traffic intensity $\rho \uparrow 1$. The second type is called the Halfin-Whitt heavy-traffic regime ([31]), in which the number of servers is allowed to grow simultaneously as the traffic intensity $\rho \uparrow 1$. In particular, in Halfin-Whitt regime, ρ scales like $1 - Bn^{-\frac{1}{2}}$ for some $B > 0$.

Results which provide error bounds on the heavy-traffic approximations are as follows. In classical heavy-traffic, this includes the work of [32, 33, 34, 35]. In the Halfin-Whitt

setting, this includes the work of [36, 37, 38, 39, 40, 41, 42, 43, 44]. In both the classical heavy-traffic and Halfin-Whitt settings, outside the case of Markovian service times, all of these results suffer from the presence of non-explicit constants, which may depend on the underlying service distribution in a very complicated and unspecified way. Furthermore, in the Halfin-Whitt setting, the limiting quantities themselves generally have no explicit representation. In both heavy-traffic settings Lyapunov function arguments have been used to yield bounds [45, 46, 47, 36], but in all cases these again suffer from the presence of non-explicit constants. Also, essentially all of the aforementioned heavy-traffic corrections require that one restrict to a specific type of heavy-traffic scaling, e.g. either classical heavy-traffic or Halfin-Whitt scaling, and do not hold universally (i.e. regardless of how one approaches heavy-traffic). Recently, some progress has been made towards developing such universal bounds for single-server systems [48] and in the presence of Markovian service times [39], but such bounds have remained elusive for general multi-server systems.

Another popular way to construct bounds for multi-server queues is to use stochastic comparison approach, which provides explicit bounds by first proving that a simpler stochastic model yields a bound on the FCFS $GI/GI/n$ queue, and then bounding this simpler system. However, bounds of this type generally do not scale correctly in the Halfin-Whitt regime (by correctly we mean scaling as $\frac{1}{1-\rho}$, as appeared in Kingman's bound). For example, by considering a modified system in which jobs are routed to individual servers cyclically (instead of FCFS), one can reduce the dynamics to those of several single-server queue, and derive the explicit bound [49, 50]

$$\mathbb{E}[W] \leq \frac{n^{-1}\sigma_S^2 + \rho(2-\rho)\sigma_A^2}{2\mathbb{E}[A]} \times \frac{1}{1-\rho} \quad ; \quad \mathbb{E}[Q] \leq \frac{n^{-1}\sigma_S^2 + \rho(2-\rho)\sigma_A^2}{2(\mathbb{E}[A])^2} \times \frac{1}{1-\rho}.$$

However, this bound on $(1-\rho)\mathbb{E}[Q]$ diverges in the Halfin-Whitt regime (as $n \to \infty$), while it is proven in [31] that $(1-\rho)\mathbb{E}[Q]$ remains uniformly bounded (independent of n), making this bound scale "incorrectly" under the Halfin-Whitt regime. Bridging this divide

has been an open question for some time [50], and relates fundamentally to the question of how a general $GI/GI/n$ queue (possibly under the Halfin-Whitt scaling) relates to the corresponding single-server queue, which has the same inter-arrival distribution, but in which service times are scaled down by n. Although this connection has been formalized in the setting of classical heavy traffic [51, 35], for general $GI/GI/n$ queues (and e.g. under the Halfin-Whitt scaling) effective upper bounds remain elusive [52, 53, 50]. Other stochastic comparison approaches were taken in [54, 55, 56, 57], although the bounds of [54] are weaker than (2.3), and the results of [55, 56, 57] (although decisive for understanding which moments of the waiting time distribution are finite) have not yielded effective upper bounds which scale correctly in the Halfin-Whitt regime.

1.3 Heavy tails and large deviations in Halfin-Whitt regime

A key insight from modern queueing theory is that when inter-arrival or service times have a heavy tail (i.e. the tail of the probability distribution does not decay exponentially), the underlying system behaves qualitatively different, e.g. it may exhibit long-range dependencies over time, and have a higher probability of rare events [58]. As several studies have empirically verified the heavy tail phenomena in applications relevant to the Halfin-Whitt scaling [7, 59], it is important to understand how the presence of heavy tails changes the performance of multi-server queues in the Halfin-Whitt scaling regime. Although there is a vast literature on parallel server queues with heavy-tailed inter-arrival and/or service times (which we make no attempt to survey here, instead referring the reader to [60]), it seems that surprisingly, very little is known about how such systems behave qualitatively in the Halfin-Whitt regime.

We now survey what is known in this setting. The results of [61, 62] imply that when inter-arrival times have finite second moment (i.e. satisfy a classical central limit theorem) and service times have finite mean (but may have infinite $1+\epsilon$ moment for some $\epsilon \in (0,1)$), the associated sequence of transient queue-length processes, normalized by $n^{\frac{1}{2}}$, converges

weakly (over compact time sets) to a non-trivial limiting process (if the system is initialized appropriately), described implicitly as the solution to a certain stochastic convolution equation.

[1] considers the case in which inter-arrival times have (asymptotically) a so-called pure Pareto tail with index $\alpha \in (1,2)$, i.e. $\lim_{x\to\infty} \frac{\mathbb{P}(A>x)}{x^{\alpha}} = C$ for some $\alpha \in (1,2)$ and $C \in (0,\infty)$, and service times are deterministic. A certain modification of the Halfin-Whitt scaling regime was considered, under which traffic intensity ρ of the sequence of $GI/D/n$ queues scales like $1 - Bn^{-\frac{1}{\alpha}}$ for some strictly positive excess parameter B. The result of [1] implies that the sequence of steady-state queue-length distributions, normalized by $n^{\frac{1}{\alpha}}$ converges to $\hat{W}$, which is the supremum of a so-called α-stable random walk with drift $-B$. Namely, for $\alpha < 2$, $n^{\frac{1}{2}}$ is no longer the correct scaling. This insight is quite interesting, although we note the important fact that Reed's results are restricted to the case of deterministic service times.

Essentially all other references in the literature to queues in the Halfin-Whitt regime with heavy tails are to open questions. The question the of tightness of the associated sequence of steady-state queue length distributions, normalized by $n^{\frac{1}{2}}$, is similarly left open when service times have infinite variance.

The presence of heavy-tailed inter-arrival or service times distributions also affects analysis on large deviations. For the case of inter-arrival times with finite second moment and service times with finite support, [46] prove that the weak limit (associated with the sequence of normalized steady-state queue lengths) has an exponential tail, with a precise exponent identified as $-\frac{2B}{c_A^2+c_S^2}$, where $c_A^2(c_S^2)$ is the squared coefficient of variation (s.c.v) of inter-arrival (service) times. Namely, they prove that under those assumptions, the associated weak limit $\hat{Q}$ satisfies $\lim_{x\to\infty} x^{-1}\log\left(\mathbb{P}\big(\hat{Q} > x\big)\right) = -\frac{2B}{c_A^2+c_S^2}$. Put another way, the probability that the limiting process exceeds a large value x behaves (roughly up to exponential order) like $\exp\big(-\frac{2B}{c_A^2+c_S^2}x\big)$. The known results for the case of exponentially distributed and H_2^* service times yields the same exponent. The stochastic comparison ap-

proach of [63] was able to prove that the same exponent yields an upper bound on the large deviations behavior of any subsequential limit of the associated sequence of normalized queue-length random variables assuming only that there exists $\epsilon > 0$ s.t. inter-arrival and service times have finite $2 + \epsilon$ moments, with equality for the case of exponentially distributed inter-arrival times. Far less is known when it comes to the large deviation behavior associated to the queueing systems with heavy-tailed inter-arrival or/and service time distribution under the Halfin-Whitt regime.

In [63], the authors note that the identified limiting large deviations exponent $-\frac{2B}{c_A^2+c_S^2}$ equals zero when either inter-arrival or service times have infinite variance, and leave as an open question identifying the correct behavior in the presence of heavy tails. This is particularly interesting to investigate as the vanishing large deviation exponent suggests that a fundamentally different behavior may arise.

1.4 Summary of contributions

For multi-server queues, our literature review reveals that no simple and explicit bounds for the steady-state queue length that scale universally as $\frac{1}{1-\rho}$ across different notions of heavy-traffic, in analogy to the celebrated Kingman's bound for single-server queues, are known. It is unclear whether such a bound is even theoretically possible, nor in what manner it would have to depend on the underlying distributions. We address this problem in Chapter 2, by developing the first simple and explicit bounds for general $GI/GI/n$ queues that scale universally as $\frac{1}{1-\rho}$ across all notions of heavy traffic, including both the classical and Halfin-Whitt scalings.

In Chapter 3, we consider the FCFS $GI/GI/n$ queue, and prove that when service times have finite $1+\epsilon$ moment for some $\epsilon > 0$ and the inter-arrival times have finite second moment, the sequence of stationary queue length distributions, normalized by $n^{\frac{1}{2}}$, is tight in the Halfin-Whitt regime, a problem previously left open. Furthermore, we develop simple and explicit bounds on the stationary queue length in that regime. For the setting where

instead the inter-arrival times have an asymptotically Pareto tail with index $\alpha \in (1,2)$, we extend recent results of [1] (who analyzed the case of deterministic service times) by proving that for general service time distributions, the sequence of stationary queue length distributions, normalized by $n^{\frac{1}{\alpha}}$, is tight (here we use the scaling of [1], which we refer to as the Halfin-Whitt-Reed scaling regime).

In Chapter 4, we further investigate the large deviation behaviors of the (any suitable subsequential) limit of the sequence of scaled stationary queue length in the presence of heavy-tailed inter-arrival and/or service times, the tightness of which were proved in Chapter 3. When service times have an asymptotically Pareto tail with index $\alpha \in (1,2)$ and inter-arrival times have finite second moment, we give upper bounds on the large deviations behavior of the limit and derive a matching lower bound when inter-arrival times are Markovian. We find that the large deviations behavior of the limit has a *sub-exponential* decay, differing fundamentally from the exponentially decaying tails known to hold in the light-tailed setting. This, in essence, resolves the question of the previously identified large deviations exponent $-\frac{2B}{c_A^2+c_S^2}$ which vanishes in the infinite-variance setting. From a practical standpoint, this insight is important, as it suggests that when service times are heavy-tailed, it is much more likely to see large queue lengths, where we successfully quantify the meaning of "much more likely". For the setting where instead the inter-arrival times have an asymptotically pure Pareto tail for some $\alpha \in (1,2)$ and service times have at least $1+\epsilon$ moment for some $\epsilon \in (0,1]$, we are again able to bound the large-deviations behavior of the limit, and find that our derived bounds do not depend on the particular service time distribution, and are in fact tight even for the case of deterministic service times.

1.5 Notations

Following notations are used throughout the thesis. Additional notations and assumptions will be introduced at the beginning of future chapters. Let us fix an arbitrary FCFS $GI/GI/n$ queue with inter-arrival distribution A and service time distribution S, and de-

note this queueing system by $\mathcal{Q}^n$. Let $\mathcal{N}_o(\mathcal{A}_o)$ denote an ordinary renewal process with renewal distribution $S(A)$, and $N_o(t)\big(A_o(t)\big)$ the corresponding counting processes. Let $\{\mathcal{N}_i, i \geq 1\}\Big(\{\mathcal{N}_{o,i}, i \geq 1\}\Big)$ denote a mutually independent collection of equilibrium (ordinary) renewal processes with renewal distribution S; $\mathcal{A}$ an independent equilibrium renewal process with renewal distribution A; and $\{N_i(t), i \geq 1\}\Big(\{N_{o,i}(t), i \geq 1\}\Big)$, $A(t)$ the corresponding counting processes. Here we recall that an equilibrium renewal process is one in which the first renewal interval is distributed as the equilibrium distribution associated with X, i.e. letting $R(X)$ denote a r.v. such that $\mathbb{P}(R(X) > y) = \frac{1}{\mathbb{E}[X]} \int_y^\infty \mathbb{P}(X > z)dz$, the first renewal interval is distributed as $R(X)$. Also, let $\mu_A(\mu_S)$ denote $\frac{1}{\mathbb{E}[A]}(\frac{1}{\mathbb{E}[S]})$; $\sigma_A(\sigma_S)$ denote $\big(\mathbb{E}[A^2] - \mathbb{E}^2[A]\big)^{\frac{1}{2}}\Big(\big(\mathbb{E}[S^2] - \mathbb{E}^2[S]\big)^{\frac{1}{2}}\Big)$; $c_A(c_S)$ denote $\mu_A\sigma_A(\mu_S\sigma_S)$; and $Var[A](Var[S])$ denote $\sigma_A^2(\sigma_S^2)$. Also, let $\{A_i, i \geq 1\}$ ($\{S_i, i \geq 1\}$) denote the sequence of inter-event times in $\mathcal{A}_o(\mathcal{N}_o)$. Let us evaluate all empty summations to zero, and all empty products to unity; and as a convention take $\frac{1}{\infty} = 0$ and $\frac{1}{0} = \infty$.

CHAPTER 2

SIMPLE AND EXPLICIT BOUNDS FOR MULTI-SERVER QUEUES WITH UNIVERSAL $\frac{1}{1-\rho}$ SCALING

2.1 Introduction.

The multi-server queue with independent and identically distributed (i.i.d.) inter-arrival and service times, and first-come-first-serve (FCFS) service discipline, is a fundamental object of study in Operations Research and Applied Probability. Its study was originally motivated by the design of telecommunication networks in the early 20th century, and pioneered by engineers such as Erlang [64]. Since that time the model has found many additional applications across a wide range of domains [65]. In the pioneering work of Erlang, it was realized that when both inter-arrival and service times are exponentially distributed (i.e. $M/M/n$) all relevant quantities can be computed in (essentially) closed form. These insights were extended to the case of multi-server queues with exponentially distributed service times and general inter-arrival times (i.e. $GI/M/n$) in [66] by considering the relevant embedded Markov chain. For the setting of non-Markovian service times, early progress was made in the analysis of single-server queues. In the early 20th century Pollaczek and Khintchine derived an explicit formula for the expected number in queue in such systems when inter-arrival times are Markovian (i.e. $M/GI/1$). Lindley, Spitzer, and others developed the theoretical foundation for analyzing general single-server models (i.e. $GI/GI/1$) as the suprema of one-dimensional random walks with i.i.d. increments.

Another key result for $GI/GI/1$ queues came in the seminal 1962 paper of John Kingman [30], in which a simple and explicit upper bound was given for the steady-state expected waiting time $\mathbb{E}[W]$. We note that by Little's law, any such bound for $\mathbb{E}[W]$ yields a corresponding bound for the steady-state expected number of jobs waiting in queue (ex-

cluding those jobs in service) $\mathbb{E}[Q]$. This bound, now referred to as Kingman's bound, states that

$$\mathbb{E}[W] \leq \frac{\sigma_A^2 + \sigma_S^2}{2\mathbb{E}[A]} \times \frac{1}{1-\rho} \quad , \quad \mathbb{E}[Q] \leq \frac{\sigma_A^2 + \sigma_S^2}{2(\mathbb{E}[A])^2} \times \frac{1}{1-\rho}, \tag{2.1}$$

with $\sigma_A^2(\sigma_S^2)$ the variance of the inter-arrival (service) time distribution, $\mathbb{E}[A]$ the mean inter-arrival time, and ρ the **traffic intensity**, defined (for the $GI/GI/1$ queue) as the ratio of the mean service time $\mathbb{E}[S]$ to the mean inter-arrival time $\mathbb{E}[A]$. In the same paper, Kingman established that (under appropriate technical conditions) this bound becomes tight as $\rho \uparrow 1$, i.e. if one considers a sequence of $GI/GI/1$ queues in heavy-traffic. As most performance metrics of the general $GI/GI/1$ queue have no simple closed-form solution, this combination of **simplicity, accuracy, and scalability** has made Kingman's bound very attractive over the years, from the perspective of both real queueing applications and as a theoretical tool.

Unfortunately, our understanding of $GI/GI/n$ queues has lagged behind. The primary reason for this stagnation has been known since the 1950's, as can be seen in the following quotation of Kendall [67], reflecting on his own work and that of others on single-server queues in 1951:

"I have had very little to say about the much more difficult problems associated with the compound queue formed in front of n counters, where $n > 1$. A great many writers (from Erlang onwards) have discussed questions of this kind, and particular reference should be made to the series of papers by Pollaczek. . .If the assumption of a negative-exponential distribution of service times is dropped, then the problem becomes instantly very much more difficult, and it is possible to give a rather interesting reason for this. With perfectly general input and service time distributions the only regeneration points for the process are . . . epochs at which a new customer arrives and finds the counter free . . . and . . . it is no

longer possible to bridge the gap between one regeneration point and the next in any simple way."

2.1.1 Multi-server queues in the Halfin-Whitt heavy-traffic regime.

More recent research on $GI/GI/n$ queues has brought to light another barrier to progress, especially on the **simplicity** front. Namely, if one simultaneously lets the number of servers n grow while letting the traffic intensity ρ (for $GI/GI/n$ queues defined as $\frac{\mathbb{E}[S]}{n\mathbb{E}[A]}$) approach one, the limiting dynamics of the multi-server queue may have a very complex form. This phenomena has been well-studied in the so-called Halfin-Whitt heavy-traffic regime, in which one considers a sequence of $GI/GI/n$ queues along which ρ scales like $1 - Bn^{-\frac{1}{2}}$ for some fixed $B > 0$. In this case, the relevant limiting dynamics are captured by complex infinite-dimensional (i.e. measure-valued) processes [68], which are quite complex even when service times are restricted to having finite support [46] or being of phase-type [36], bringing into doubt the prospects of a "simple formula" such as Kingman's bound here.

Indeed, all known simple and explicit bounds for general multi-server queues scale incorrectly in the Halfin-Whitt regime. With that in mind, we now review several relevant settings in which various bounds for multi-server queues are known. These will all generally suffer from at least one of the following problems:

- Scales correctly only when the number of servers is held fixed as $\rho \uparrow 1$;
- Holds only asymptotically;
- Involves non-explicit constants;
- Holds only for restrictive classes of service time distributions;
- Holds only under a specific type of heavy-traffic scaling.

2.1.2 Multi-server queues in the classical heavy-traffic regime

If instead of a Halfin-Whitt type scaling, one first fixes the number of servers, and then considers a sequence of $GI/GI/n$ queues along which $\rho \uparrow 1$, much more is known. In this setting, known as the classical heavy-traffic regime for multi-server queues, it is known that a Kingman-type result holds asymptotically (as $\rho \uparrow 1$) for $\mathbb{E}[W]$. In particular, under appropriate technical conditions,

$$\lim_{\rho\uparrow 1}\Big((1-\rho)\mathbb{E}[W]\Big) = \frac{n^{-2}\sigma_S^2+\sigma_A^2}{2\mathbb{E}[A]} \quad , \quad \lim_{\rho\uparrow 1}\Big((1-\rho)\mathbb{E}[Q]\Big) = \frac{n^{-2}\sigma_S^2+\sigma_A^2}{2(\mathbb{E}[A])^2}. \tag{2.2}$$

Such a result was first conjectured in [69], and later proven in [70], where we again note that in contrast to the single-server case, there is no simple explicit bound for any given system, and the result only holds in an asymptotic sense as $\rho \uparrow 1$. The underlying reason why such a simple asymptotic result holds is that under this scaling, the relevant limiting dynamics correspond to a (simple) 1-dimensional reflected Brownian motion; alternatively, the multi-server system behaves like a sped-up single-server queue [71, 72, 51, 70].

Although (2.2) holds only asymptotically, there are many non-asymptotic (i.e. quantitative) bounds available in the literature for $GI/GI/n$ queues. These are generally of two fundamental types.

2.1.3 Error bounds on heavy-traffic approximations

First, there are results which provide error bounds on the aforementioned heavy-traffic approximations. In classical heavy-traffic, this includes the work of [32, 33, 34, 35]. In the Halfin-Whitt setting, this includes the work of [36, 37, 38, 39, 40, 41, 42, 43, 44]. In both the classical heavy-traffic and Halfin-Whitt settings, outside the case of Markovian service times, all of these results suffer from the presence of non-explicit constants, which may depend on the underlying service time distribution in a very complicated and unspecified way. Furthermore, in the Halfin-Whitt setting, the limiting quantities them-

selves generally have no explicit representation. In both heavy-traffic settings Lyapunov function arguments have been used to yield bounds [45, 46, 47, 36], but in all cases, these again suffer from the presence of non-explicit constants. In addition, the subset of the aforementioned results applicable in the Halfin-Whitt setting generally requires that the service time distribution comes from a restrictive class (such as phase-type), and/or makes other technical assumptions (e.g. the presence of a strictly positive abandonment rate). We note that even though phase-type distributions can in principle approximate any distribution arbitrarily well, quantitative results in the Halfin-Whitt scaling proven for phase-type distributions cannot be easily extended by approximating a general distribution in this way, since there may be complex interactions between the error in the phase-type approximation, the heavy-traffic error bounds, and the discrepancy between the limiting dynamics arising from a phase-type service time distribution and a general service time distribution. The same is true if one tries to take a result which holds in this setting with strictly positive abandonments, and attempts to let the abandonment rate go to zero. In addition, essentially all of the aforementioned heavy-traffic corrections require that one restrict to a specific type of heavy-traffic scaling, e.g. either classical heavy-traffic or Halfin-Whitt scaling, and do not hold universally (i.e. regardless of how one approaches heavy-traffic). Recently, some progress has been made towards developing such universal bounds for single-server systems [48] and in the presence of Markovian service times [39], but such bounds have remained elusive for general multi-server systems.

2.1.4 Stochastic comparison results

Second, there are results which provide explicit bounds by first proving that a simpler stochastic model yields a bound on the FCFS $GI/GI/n$ queues, and then bounding this simpler system. By considering a modified system in which jobs are routed to individual servers cyclically (instead of FCFS), one can reduce the dynamics to those of several single-

server queue, and derive the explicit bound [49, 50]

$$\mathbb{E}[W] \leq \frac{n^{-1}\sigma_S^2 + \rho(2-\rho)\sigma_A^2}{2\mathbb{E}[A]} \times \frac{1}{1-\rho} \quad ; \quad \mathbb{E}[Q] \leq \frac{n^{-1}\sigma_S^2 + \rho(2-\rho)\sigma_A^2}{2(\mathbb{E}[A])^2} \times \frac{1}{1-\rho}. \quad (2.3)$$

Comparing (2.3) to (2.2), one sees the crucial difference: the price one pays to get an explicit result holding universally is to replace the term $n^{-2}\sigma_S^2$ by $n^{-1}\sigma_S^2$. In classical heavy-traffic, as $\rho \uparrow 1$, this leads to a multiplicative error which grows with n, but remains bounded as $\rho \uparrow 1$ for any fixed n. However, in the Halfin-Whitt type scaling, this leads to bounds which scale in a fundamentally incorrect nature. Indeed, it is proven in [31] that $(1-\rho)\mathbb{E}[Q]$ remains uniformly bounded (independent of n) when considering a sequence of $M/M/n$ queues in the Halfin-Whitt scaling, while the bound for $(1-\rho)\mathbb{E}[Q]$ given by (2.3) would diverge (growing linearly in n) along such a sequence. Bridging this divide has been an open question for some time [50], and relates fundamentally to the question of how a general $GI/GI/n$ queue (possibly under the Halfin-Whitt scaling) relates to the corresponding single-server queue, which has the same inter-arrival distribution, but in which service times are scaled down by n. Although this connection has been formalized in the setting of classical heavy traffic [51, 35], for general $GI/GI/n$ queues (and e.g. under the Halfin-Whitt scaling) effective upper bounds remain elusive [52, 53, 50]. Other stochastic comparison approaches were taken in [54, 55, 56, 57], although the bounds of [54] are weaker than (2.3), and the results of [55, 56, 57] (although decisive for understanding which moments of the waiting time distribution are finite) have not yielded effective upper bounds which scale correctly in the Halfin-Whitt regime.

There is also a literature on conjectural results (i.e. unproven) as regards stochastic comparison results for multi-server queues. Several of these revolve around using convexity notions to bound the waiting time in a $M/GI/n$ queue from above (below) by the waiting time when the service time distribution is either deterministic or a certain extremal mixture of exponentials [73, 74, 50], where we note that such comparisons are known to

hold in the single-server setting [75, 76]. Although there has been some success in proving considerably weaker forms of these conjectures in the multi-server setting [77], these results generally involve correction terms which render them ineffective in the Halfin-Whitt scaling. In addition to conjectural results along these lines, [73] also provides explicit examples of pairs of $M/GI/n$ queues for which the inter-arrival and service time distributions have the same first two moments, yet $\mathbb{E}[W]$ differs considerably, suggesting that even if some sort of simple explicit bounds were to exist for general multi-server queues, there would be no hope of having them be "asymptotically tight" in any sort of universal sense. Another series of stochastic-comparison related conjectures revolves around the system in which the servers are partitioned into clusters, with one cluster for each different value a job's service time (or exponential rate of service in the case that service times are a finite mixture of exponentials) can take, and all jobs with that service time being sent to the appropriate dedicated cluster (at which they are processed according to FCFS) [78, 79]. It remains an open question when such a system will yield an upper bound on the original multi-server system, since although the modified system is not fully taking advantage of resource pooling, it may avoid situations in which a job with a large service time blocks many jobs with smaller service times [80]. We note that several authors have pointed out that stochastic comparison results can be quite subtle, and the literature contains several "proofs" of "obvious comparisons" which were later found to be incorrect [81, 50]. We also note that most of the stochastic comparison literature focuses on quantities such as waiting time and queue length, as opposed to the steady-state probability of delay (s.s.p.d., i.e. steady-state probability that all servers are busy), for which one of the few explicit results is that of [82], who proved that in a $GI/GI/n$ queue, the s.s.p.d. is always between $\rho - (n-1)(1-\rho)$ and ρ (a bound which is only effective when ρ is very close to 1, i.e. outside the Halfin-Whitt regime).

Two stochastic comparison results of [63] and [83] are especially pertinent to our own investigations, in which stochastic comparison arguments were used to bound the num-

ber of jobs waiting in queue in a $GI/GI/n$ queue by the supremum of a one-dimensional random walk involving the original arrival process and a pooled renewal process (with renewal intervals corresponding to the service time distribution). The weak limit of this supremum (under the Halfin-Whitt scaling) was analyzed in [63], and used to show that the steady-state queue length scales like $n^{\frac{1}{2}}$, and to bound the large deviations behavior of the associated limiting process. These results were extended by considering a closely related stochastic comparison argument in [83], which was able to yield bounds on the s.s.p.d. when framed as a double limit within the Halfin-Whitt regime, i.e. if for each fixed $B > 0$ one considers a sequence of bounding processes along which ρ scales like $1 - Bn^{-\frac{1}{2}}$, and then analyzes the limit of this sequence (for each fixed B, letting $n \to \infty$) along a sequence of B which approach either 0 or ∞. Although providing several qualitative insights into how various quantities behave asymptotically in the Halfin-Whitt regime, these results did not provide any explicit bounds for any fixed $GI/GI/n$ system, as the aforementioned supremum was only analyzed asymptotically (within the Halfin-Whitt regime), and even then in many regards the associated weak limit was itself only analyzed as certain parameters approached either 0 or ∞.

2.1.5 Why scaling as $\frac{1}{1-\rho}$?

In the above discussion, we several times referenced the fact that certain bounds "scaled as $\frac{1}{1-\rho}$", or "did not scale correctly" because they did not scale as $\frac{1}{1-\rho}$. It is of course reasonable to ask why, and in what precise sense, $\frac{1}{1-\rho}$ should be the bar. There are at least two fundamental justifications here. First, it follows from well-known results for the M/M/n queue [31] that for any fixed $B > 0$, there exist $L_B, U_B \in (0, \infty)$ such that any $M/M/n$ queue for which $\rho \in (1 - Bn^{-\frac{1}{2}}, 1)$ satisfies $L_B \times \frac{1}{1-\rho} \leq \mathbb{E}[Q] \leq U_B \times \frac{1}{1-\rho}$. Thus, in a fairly general sense, this is the correct scaling for the $M/M/n$ queue in heavy-traffic. Second, this is the correct scaling in both classical heavy-traffic and the Halfin-Whitt scaling. In particular, if $\{Q_n, n \geq 1\}$ is a sequence of random variables corresponding to the

steady-state queue-lengths of a sequence of multi-server queues being scaled according to either classical heavy-traffic, Halfin-Whitt heavy-traffic, or even non-degenerate slow-down heavy-traffic [84], then letting ρ_n denote the traffic intensity of the nth queue in the sequence, one has (under appropriate technical conditions) that $\{(1-\rho_n)Q_n, n \geq 1\}$ is tight and has a non-degenerate limit [70, 63, 68, 84]. Indeed, the $\frac{1}{1-\rho}$ scaling is a guiding meta-principle throughout the entire literature on multi-server queues. We note that different works often present / realize this phenomena from slightly different angles, e.g. one may get a slightly different result if analyzing waiting times (as opposed to queue lengths) as one must apply Little's law to appropriately "translate" the $\frac{1}{1-\rho}$ scaling, and many papers may formally prove a weak-convergence type result without formally proving that the corresponding sequence of expected values scales in the same way.

Of course, it is possible to come up with sequences of queues where $\frac{1}{1-\rho}$ is not the appropriate scaling. For example, it follows from well-known results for the M/M/n queue [31] that in that special setting, $\mathbb{E}[Q] = P(Q > 0) \times \frac{\rho}{1-\rho}$, where $P(Q > 0)$ is the s.s.p.d. Thus as either ρ or $P(Q > 0)$ approaches 0, $\frac{1}{1-\rho}$ will no longer be asymptotically correct. For example, suppose one has an $M/M/n$ queue with n large and $\rho = 1 - n^{-\frac{1}{4}}$. In that case, it follows from [31] that $P(Q > 0) \to 0$ as $n \to \infty$, and hence $\frac{1}{1-\rho}$ would significantly overestimate the expected queue length. Also, one can always construct pathological service time distributions which exhibit atypical behavior. For example, one could consider a sequence of multi-server queues along which $\mathbb{E}[S]$ is held fixed, while $\mathbb{E}[S^2]$ and n both diverge, and $\rho \uparrow 1$. By linking the higher moments of the service time distribution to the number of servers, one can also induce certain pathologies which would render the $\frac{1}{1-\rho}$ scaling incorrectly. None-the-less, as explained above, the literature certainly supports the general rule-of-thumb that for a multi-server queue in heavy-traffic, Q should generally scale as $\frac{1}{1-\rho}$.

2.1.6 Other approaches to analyzing queueing systems

There is a vast literature on numerical approaches to understanding $GI/GI/n$ queues, including simulation [85, 86], the computation and subsequent inversion of transforms [87], matrix-analytic methods [88], numerical analysis of diffusions [89], convex optimization [90], and robust optimization [91]. These methods have their own pros and cons in different settings, but generally have a different aim than developing simple and explicit formulas useful in the creation of easy heuristics, which will be the subject of our own investigations. As such, we will not discuss those methods and the associated literature in any depth, nor will we discuss the vast literature on more complicated queueing models (e.g. queueing networks, systems with heavy tails, etc.), or heuristic approaches which are not rigorously justified. We do note that to our knowledge, all of the simple and explicit analytical bounds derived from these alternative approaches (which are rigorously justified), e.g. the explicit analogue of Kingman's bound for multi-server queues derived in [91], suffer from many of the same shortcomings mentioned earlier (e.g. incorrect scaling in the Halfin-Whitt regime). We also note that there have been many surveys written on the different approaches to queueing theory over the years, and refer the reader to [92, 93, 94].

2.1.7 Main contribution

Summary of state-of-the-art

To summarize the above literature review, the state-of-the-art for $GI/GI/n$ queues when service times are neither deterministic nor exponentially distributed is as follows. In spite of thousands of papers devoted to the theoretical analysis of multi-server queues, no simple and explicit bounds for the steady-state queue length that scale universally as $\frac{1}{1-\rho}$ across different notions of heavy-traffic, in analogy to the celebrated Kingman's bound for single-server queues, are known. Furthermore, it is unclear whether such a bound is even theoretically possible, nor in what manner it would have to depend on the underlying distributions.

Our contribution

In this chapter, we develop the first simple and explicit bounds for general $GI/GI/n$ queues that scale universally as $\frac{1}{1-\rho}$ across all notions of heavy traffic, including both the classical and Halfin-Whitt scalings.

2.1.8 Chapter outline

The remainder of the chapter proceeds as follows. In Section 2.2, we state our main results, along with several interesting corollaries and applications. In Section 2.3, we review the stochastic comparison results of [63] and [83], upon which our analysis will build. In Section 2.4, we provide very general conditional bounds on the supremum which arises in [63, 83], where these conditional bounds are of the form (for example) "If the moments of certain processes satisfy ..., then the supremum of interest satisfies ...". In Section 2.5, we provide an in-depth analysis of the processes arising in the aforementioned conditional bounds, notably certain pooled renewal processes, under the assumption that inter-arrival and service times have finite rth moment for some $r > 2$. Combined with our previous conditional results, we then complete the proof of our main results. We provide a summary of our results, concluding remarks, and some directions for future research in Section 2.6. Finally, we include a technical appendix in Section 2.7, which contains several technical arguments from throughout the chapter.

2.2 Main Results

2.2.1 Additional Notations

In addition to notations introduced in Section 1.5, for our results involving steady-state queue lengths, we will generally require that for any given initial condition, the total number of jobs in $\mathcal{Q}^n$ (number in service + number waiting in queue) converges in distribution (as time goes to infinity, independent of the particular initial condition) to a steady-state

r.v. $Q^n(\infty)$. As a shorthand, we will denote this assumption by saying "$Q^n(\infty)$ exists". We will adopt a parallel convention when talking about the steady-state waiting time of an arriving job. Namely, for our results on waiting times, we will generally require that for any given initial condition, the distribution of the waiting time (in queue, not counting time in service) of the kth arrival to the system converges in distribution (as $k \to \infty$, independent of the particular initial condition) to a steady-state r.v. $W^n(\infty)$. As a shorthand, we will denote this assumption by saying "$W^n(\infty)$ exists". Also, supposing that $Q^n(\infty)$ exists, let $L^n(\infty)$ denote the steady-state number of jobs waiting in queue, i.e. $L^n(\infty)$ is distributed as $\max\big(0, Q^n(\infty) - n\big)$. For $k \geq 1$, let $\rho_k \triangleq \frac{\mu_A}{k\mu_S}$. Note that for any $GI/GI/n$ queue, one can always rescale both the service and inter-arrival times so that $\mathbb{E}[S] = \mu_S = 1$, without changing either ρ or the distribution of $Q^n(\infty)$. As doing so will simplify (notationally) several arguments and statements, sometimes we impose the additional assumption that $\mathbb{E}[S] = \mu_S = 1$, and will point out whenever this is the case.

2.2.2 Main results

Our main results are the following novel, explicit bounds for general multi-server queues whose inter-arrival and service time distributions have finite $2 + \epsilon$ moments, which scale universally as $\frac{1}{1-\rho}$. Our bounds depend only on a single moment of the service and inter-arrival time distributions, and are the first such explicit bounds for general multi-server queues.

Theorem 2.1. *Suppose that for a $GI/GI/n$ queue with inter-arrival distribution A and service time distribution S, there exists $r > 2$ s.t. $\mathbb{E}[S^r] < \infty, E[A^r] < \infty$. Suppose also that $0 < \mu_A < n\mu_S < \infty$, and that $Q^n(\infty)$ exists. Then for all $x > 0$, $\mathbb{P}\big(L^n(\infty) \geq \frac{x}{1-\rho_n}\big)$ is at most*

$$\left(\mathbb{E}\big[(S\mu_S)^r\big]\mathbb{E}\big[(A\mu_A)^r\big]\right)^3 \left(10^{120} r^{32} (r-2)^{-12}\right)^r x^{-\frac{r}{2}};$$

and the steady-state probability of delay (s.s.p.d.), $\mathbb{P}\big(Q^n(\infty) \geq n\big)$*, is at most*

$$\left(\mathbb{E}\big[(S\mu_S)^r\big]\mathbb{E}\big[(A\mu_A)^r\big]\right)^3\left(10^{120}r^{32}(r-2)^{-12}\right)^r\left(n(1-\rho_n)^2\right)^{-\frac{r}{2}}.$$

2.2.3 Further implications of our main results

We now introduce several direct implications of our main results, further emphasizing their utility. In all cases these results follow directly from our main results, straightforward algebra/calculus, and Little's Law.

We first state several implications for the mean steady-state waiting time and number in queue.

Corollary 2.1. *Under the same assumptions as Theorem 2.1, and supposing in addition that* $W^n(\infty)$ *exists, then*

$$\mathbb{E}\big[L^n(\infty)\big] \leq \left(\mathbb{E}\big[(S\mu_S)^r\big]\mathbb{E}\big[(A\mu_A)^r\big]\right)^3\left(10^{121}r^{33}(r-2)^{-13}\right)^r \times \frac{1}{1-\rho_n};$$

$$\mathbb{E}\big[W^n(\infty)\big] \leq \mathbb{E}[A] \times \left(\mathbb{E}\big[(S\mu_S)^r\big]\mathbb{E}\big[(A\mu_A)^r\big]\right)^3\left(10^{121}r^{33}(r-2)^{-13}\right)^r \times \frac{1}{1-\rho_n}.$$

We now state additional implications for higher moments. We note in addition to being the first such bounds for multi-server queues which scale universally as $\frac{1}{1-\rho}$, it seems that our bounds even shed new light on the single-server queue, for which the past literature on explicit bounds for higher moments seems to be largely restricted to non-explicit recursive formulas [95].

Corollary 2.2. *Under the same assumptions as Corollary 2.1, for all* $z \in [1, \frac{r}{2})$*,*

$$\mathbb{E}\big[\big(L^n(\infty)\big)^z\big] \leq \left(\mathbb{E}\big[(S\mu_S)^r\big]\mathbb{E}\big[(A\mu_A)^r\big]\right)^3\left(10^{121}r^{33}(r-2)^{-12}(r-2z)^{-1}\right)^r \times \big(\frac{1}{1-\rho_n}\big)^z.$$

We note that other moment relations for various additional quantities could also be

obtained using e.g. the generalized Little's Law [96] and distributional Little's Law [97], although we do not pursue that here.

We now state several implications for queues under the Halfin-Whitt scaling. We state two types of results. First, we state results for general queues which are not explicitly in the "Halfin-Whitt" scaling as traditionally defined (i.e. for an appropriate sequence of queues), but which satisfy as an inequality the traffic intensity condition of the Halfin-Whitt regime.

Corollary 2.3. *Under the same assumptions as Theorem 2.1, and supposing in addition that $\rho_n \leq 1 - Bn^{-\frac{1}{2}}$ for some $B > 0$, the following holds. For all $x > 0$, $\mathbb{P}\big(L^n(\infty) \geq xn^{\frac{1}{2}}\big)$ is at most*

$$\left(\mathbb{E}\big[(S\mu_S)^r\big]\mathbb{E}\big[(A\mu_A)^r\big]\right)^3 \left(10^{120} r^{32} (r-2)^{-12}\right)^r B^{-\frac{r}{2}} x^{-\frac{r}{2}};$$

and the steady-state probability of delay (s.s.p.d.), $\mathbb{P}\big(Q^n(\infty) \geq n\big)$, is at most

$$\left(\mathbb{E}\big[(S\mu_S)^r\big]\mathbb{E}\big[(A\mu_A)^r\big]\right)^3 \left(10^{120} r^{32} (r-2)^{-12}\right)^r B^{-r}.$$

Next, we state the corresponding results for a sequence of queues explicitly in the "Halfin-Whitt" scaling as traditionally defined. Namely, let us fix non-negative unit mean r.v.s $\hat{A}$ and $\hat{S}$, and a real number $B > 0$. For $n > B^2$, let $\hat{Q}_B^n(\infty)$ denote a r.v. distributed as the steady-state number in system (number in service + number waiting in queue) in the $GI/GI/n$ queue with inter-arrival distribution $\frac{\hat{A}}{n - Bn^{\frac{1}{2}}}$ and service time distribution $\hat{S}$ (supposing that all relevant steady-state quantities exist). Let $\hat{L}_B^n(\infty)$ denote a r.v. distributed as the corresponding steady-state number waiting in queue. Then our results imply the following.

Corollary 2.4. *Suppose that for some $r > 2$, it holds that $\mathbb{E}[\hat{A}^r] < \infty, \mathbb{E}[\hat{S}^r] < \infty$, and that $n > B^2$. Then for all $x > 0$, $\mathbb{P}\big(n^{-\frac{1}{2}} \hat{L}_B^n(\infty) \geq x\big)$ is at most*

$$\left(\mathbb{E}\big[\hat{S}^r\big]\mathbb{E}\big[\hat{A}^r\big]\right)^3 \left(10^{120} r^{32} (r-2)^{-12}\right)^r B^{-\frac{r}{2}} x^{-\frac{r}{2}};$$

for all $z \in [1, \frac{r}{2})$,

$$\mathbb{E}\left[\left(n^{-\frac{1}{2}}\hat{L}_B^n(\infty)\right)^z\right] \leq \left(\mathbb{E}\big[\hat{S}^r\big]\mathbb{E}\big[\hat{A}^r\big]\right)^3 \left(10^{121} r^{33} (r-2)^{-12} (r-2z)^{-1}\right)^r B^{-\frac{r}{2}};$$

and the steady-state probability of delay (s.s.p.d.), $\mathbb{P}\big(\hat{Q}_B^n(\infty) \geq n\big)$*, is at most*

$$\left(\mathbb{E}\big[\hat{S}^r\big]\mathbb{E}\big[\hat{A}^r\big]\right)^3 \left(10^{120} r^{32} (r-2)^{-12}\right)^r B^{-r}.$$

These results give explicit and universal bounds on the steady-state queue length, for queues in the Halfin-Whitt regime, in terms of only a single moment of $\hat{A}$ and $\hat{S}$, and the excess parameter B. These results are the first of their kind for queues in this regime, for which (as discussed earlier) all previous explicit results were known only for the case of Markovian service times. These results also have important implications for the s.s.p.d. Namely, they give simple, explicit, non-asymptotic bounds on how the s.s.p.d. decays with B. Indeed, although the Halfin-Whitt regime is (in a sense) defined by the s.s.p.d. having a non-trivial value in $(0, 1)$ even for very large numbers of servers, no simple, explicit, non-asymptotic bounds on this quantity were previously known. To further illustrate the result, let us give an even more concrete version of Corollary 2.4.

Corollary 2.5. *Suppose that, in addition to the assumptions of Corollary 2.4, it holds that* $\mathbb{E}[\hat{S}^3] \leq 10^3, \mathbb{E}[\hat{A}^3] \leq 10^3$*. Then the steady-state probability of delay (s.s.p.d.),* $\mathbb{P}\big(\hat{Q}_B^n(\infty) \geq n\big)$*, is at most*

$$10^{500} B^{-3}.$$

Thus (for example) under only a third-moment assumption, the s.s.p.d. in the Halfin-Whitt regime decays as B^{-3} for large B, independent of the number of servers.

Let us now briefly take a moment to address the proverbial "elephant in the room" - namely, the massive prefactors in these results. One important point is that in all proofs, simplicity was opted for over tightening these constants. Thus, presumably a more careful

analysis using essentially the same exact ideas would lead to a significantly reduced prefactor. Furthermore, we view our results as a significant "proof-of-concept" as regards simple and explicit bounds for multi-server queues, and believe that future work, building on our own, will ultimately lead to the formulation of more practical bounds.

On a related note, the results of [83] imply that, in the Halfin-Whitt regime, if one is willing to settle for a purely asymptotic result, then for large B the s.s.p.d. actually has a Gaussian decay (in B), where we note that such a decay has also been observed for alternative service disciplines (i.e. not FCFS) in [79] (under additional technical assumptions). The results of [63] similarly imply that, again if one is willing to settle for a purely asymptotic and non-explicit result, then for large x the probability that the rescaled queue-length exceeds x should have an exponential decay (in x). Also, the results of [55, 56] imply the existence of more finite moments (for e.g. the queue length) than is implied by our own results. In all cases, bridging these gaps remain interesting open questions, and we refer the reader to [47] for some further relevant discussion as regards bridging asymptotic and non-asymptotic results in the Halfin-Whitt regime. Of cousre, as mentioned earlier, acheiving a simple and explicit bound which not only scales correctly, but is actually exact in heavy-traffic (a la Kingman's bound) may actually be impossible in the Halfin-Whitt regime, as the underlying limit processes seem to be inherently complicated.

2.3 Review of upper bounds from [63]

In [63], the authors proves that $Q^n(\infty)$ can be bounded from above (in distribution) by the supremum of a relatively simple one-dimensional random walk. We note that although to simplify notations the authors of [63] imposed the restriction that $\mathbb{P}(A = 0) = \mathbb{P}(S = 0) = 0$ (to preclude having to deal with simultaneous events), this restriction is unnecessary and the proofs of [63] can be trivially modified to accomodate this setting. As such, we state the relevant stochastic-comparison result of [63] here without that unnecessary assumption.

Theorem 2.2 ([63]). *Suppose that for a $GI/GI/n$ queue with inter-arrival distribution A*

and service time distribution S, it holds that $0 < \mu_A < n\mu_S < \infty$, and that $Q^n(\infty)$ exists. Then for all $x \geq 0$,

$$\mathbb{P}\big(Q^n(\infty) - n \geq x\big) \leq \mathbb{P}\left(\sup_{t \geq 0} \left(A(t) - \sum_{i=1}^{n} N_i(t) \right) \geq x \right). \tag{2.4}$$

In [83], the author extends the framework of [63] considerably and derives analogous bounds for the s.s.p.d., for which Theorem 2.2 only provides trivial bounds. In particular, Theorem 4 of [83] implies the following bound. We note that as [83] actually states a more general result, for completeness we explicitly provide the derivation of this bound from Theorem 4 of [83] in the appendix.

Theorem 2.3 ([83])**.** *Under the same assumptions as Theorem 2.2, it holds that*

$$\mathbb{P}\big(Q^n(\infty) \geq n\big) \leq \mathbb{P}\left(\sup_{t \geq 0} \left(A(t) - \sum_{i=1}^{n - \lfloor \frac{1}{2}(n - \frac{\mu_A}{\mu_S}) \rfloor} N_i(t) \right) \geq \lfloor \frac{1}{2}(n - \frac{\mu_A}{\mu_S}) \rfloor \right). \tag{2.5}$$

2.4 Bounds for $\sup_{t \geq 0} \left(A(t) - \sum_{i=1}^{n} N_i(t) \right)$

In this section we prove explicit bounds for $\sup_{t \geq 0} \left(A(t) - \sum_{i=1}^{n} N_i(t) \right)$ under minimal assumptions. In light of Theorem's 2.2 and 2.3, such bounds will be key to deriving bounds for the $GI/GI/n$ queue. First, we introduce some additional notation, and note that many results will not be stated in terms of the number of servers n, but will instead be stated in terms of a (potentially different) number n', to allow for the application of both Theorems 2.2 and 2.3 (which require considering pooled renewal processes with different numbers of components). For $n' \geq 1$, let $\mathcal{A}_{o,n'}$ denote the ordinary renewal process with renewal distribution $\min(2\rho_{n'}, 1)A$, where we take $\mathcal{A}_{o,n'}$ independent of $\{\mathcal{N}_i, i \geq 1\}$. Also, let $A_{o,n'}(t)$ denote the corresponding counting process. Note that we may construct $\{A(t), t \geq 0\}, \{A_o(t), t \geq 0\}, \{A_{0,n'}(t), t \geq 0\}$ on the same probability space s.t. w.p.1,

$$A(t) \leq 1 + A_o(t) \leq 1 + A_{o,n'}(t) \text{ for all } t \geq 0, \tag{2.6}$$

the final inequality following from the fact that since the renewal distribution of $\mathcal{A}_{o,n'}$ is a constant (at most one) multiple of the renewal distribution of $\mathcal{A}$, both may be constructed on the same probability space s.t. w.p.1 $A_{o,n'}(t) = A_o\big(\frac{t}{\min(2\rho_{n'},1)}\big) \geq A_o(t)$. For $n' \geq 1$, let $\mu_{A,n'} \stackrel{\Delta}{=} \frac{\mu_A}{\min(2\rho_{n'},1)}$, where we note that $\mu_{A,n'} = \max(\frac{1}{2}n'\mu_S, \mu_A)$. Also, let $\{A_{i,n'}, i \geq 1\}$ denote the sequence of inter-event times in $\mathcal{A}_{o,n'}$. We note that $\mathbb{E}[A_{1,n'}] = \frac{1}{\mu_{A,n'}}$, and that $\mu_A < n'\mu_S$ implies $\mu_{A,n'} < n'\mu_S$. By working with $\mathcal{A}_{o,n'}$, we will preclude certain technicalities which arise when considering queues with many servers and very low traffic intensity.

Our explicit bounds for $\sup_{t\geq 0}\big(A(t) - \sum_{i=1}^{n} N_i(t)\big)$ will be the following conditional result, which translates bounds on the moments of $|\sum_{i=1}^{n} N_i(t) - n\mu_S t|$ and $|\sum_{i=1}^{k} A_i - k\mu_A|$ into bounds on the tail of $\sup_{t\geq 0}\big(A(t) - \sum_{i=1}^{n} N_i(t)\big)$.

Theorem 2.4. *Suppose that $\mathbb{E}[S] = 1$, and that for some positive integer n' and some fixed $C_1, C_2, C_3 > 0$; $r_1 > s_1 > 1$; $r_3 > s_3 > 1$; and $r_2 > 2$, the following conditions hold:*

(i) $0 < \mu_A < n' < \infty$.

(ii) For all $t \geq 1$,

$$\mathbb{E}\big[|\sum_{i=1}^{n'} N_i(t) - n't|^{r_1}\big] \leq C_1 {n'}^{\frac{r_1}{2}} t^{s_1}.$$

(iii) For all $t \in [0,1]$,

$$\mathbb{E}\big[|\sum_{i=1}^{n'} N_i(t) - n't|^{r_2}\big] \leq C_2 \max\big(n't, (n't)^{\frac{r_2}{2}}\big).$$

(iv) For all $k \geq 1$,

$$\mathbb{E}\big[|k - \mu_A \sum_{i=1}^{k} A_i|^{r_3}\big] \leq C_3 k^{s_3}.$$

Then for all $x \geq 16$, $\mathbb{P}\left(\sup_{t\geq 0} \left(A(t) - \sum_{i=1}^{n'} N_i(t) \right) \geq x \right)$ *is at most*

$$\left(\frac{10^6 (r_1+r_2+r_3)^5}{(s_1-1)(s_3-1)(r_1-s_1)(r_3-s_3)(r_2-2)} \right)^{r_1+r_2+r_3+1} (1+C_1)(1+C_2)(1+C_3)$$
$$\times \left(n'^{\frac{r_1}{2}} (n'-\mu_{A,n'})^{-s_1} x^{-(r_1-s_1)} \right.$$
$$+ \quad n'^{\frac{r_2}{2}} (n'-\mu_{A,n'})^{-\frac{r_2}{2}} x^{-\frac{r_2}{2}}$$
$$\left. + \quad (n'-\mu_{A,n'})^{-s_3} (n')^{r_3} \mu_{A,n'}^{-(r_3-s_3)} x^{-(r_3-s_3)} \right).$$

We will prove Theorem 2.4 in several parts. First, we implement two straightforward technical simplifications. In particular, (1) we reduce the general setting to the setting in which $\rho_n \geq \frac{1}{2}$ by working with the process $\mathcal{A}_{o,n'}$, and (2) we reduce the problem to bounding two separate suprema, one for the arrival process and one for the departure process. We proceed by applying a simple union bound to the right-hand-side of (2.4), in which case we derive the following result by adding and subtracting $\frac{1}{2}(n'\mu_S + \mu_{A,n'})t = \mu_{A,n'}t + \frac{1}{2}(n'\mu_S - \mu_{A,n'})t = n'\mu_S t - \frac{1}{2}(n'\mu_S - \mu_{A,n'})t$ in (2.4), and applying (2.6).

Lemma 2.1. *Suppose that* $\mathbb{E}[S] = 1$*, and for some strictly positive integer* n'*, it holds that* $0 < \mu_A < n' < \infty$*. Then for all* $x > 2$*, it holds that* $\mathbb{P}\left(\sup_{t\geq 0} \left(A(t) - \sum_{i=1}^{n'} N_i(t) \right) \geq x \right)$ *is at most*

$$\mathbb{P}\left(\sup_{t\geq 0} \left(A_{o,n'}(t) - \mu_{A,n'}t - \frac{1}{2}(n'-\mu_{A,n'})t \right) \geq \frac{1}{2}x - 1) \right) \tag{2.7}$$
$$+ \quad \mathbb{P}\left(\sup_{t\geq 0} \left(n't - \sum_{i=1}^{n'} N_i(t) - \frac{1}{2}(n'-\mu_{A,n'})t \right) \geq \frac{1}{2}x \right). \tag{2.8}$$

The remainder of the proof of Theorem 2.4 proceeds roughly as follows.

(i) Bound the supremum of $n't - \sum_{i=1}^{n'} N_i(t)$ over sets of consecutive integers.

(ii) Bound the supremum of $n't - \sum_{i=1}^{n'} N_i(t)$ over intervals of length at most 1.

(iii) Combine (i) and (ii) to bound $\sup_{t\geq 0}\left(n't - \sum_{i=1}^{n'} N_i(t) - \frac{1}{2}(n' - \mu_{A,n'})t\right)$.

(iv) Bound the supremum of $k - \mu_{A,n'}\sum_{i=1}^{k} A_{i,n'}$ over sets of consecutive integers.

(v) Use (iv) to bound $\sup_{t\geq 0}\left(A_{o,n'}(t) - \mu_{A,n'}t - \frac{1}{2}(n' - \mu_{A,n'})t\right)$.

Before embarking on (i) - (v), we begin by reviewing a maximal inequality of Billlingsley, which will be critical for converting the moment bounds for partial sums posited in the assumptions of Theorem 2.4 into the bounds for suprema appearing in (i) - (v).

2.4.1 Review of a maximal inequality of Billingsley.

We begin by reviewing a particular maximal inequality of Billingsley. Such inequalities give general results for converting bounds on the difference between any two partial sums of a sequence of r.v.s into bound for the supremum of the partial sums of the given sequence, and are a common tool in proving tightness of stochastic processes. In particular, the following maximal inequality follows immediately from [98] Theorem 2 by setting the function they call $g(i,j)$ (defined for every pair of non-negative integers $i \leq j$) equal to $C\times(j-i+1)$ for any given $C > 0$. As the authors there note, the result also follows almost immediately from certain maximal inequalities present in an earlier edition of Billingsley's celebrated book on weak convergence [99]. We note that [98] actually proves a tighter bound, however for ease of exposition we present the following simpler bound which follows directly from [98]. For completeness, we include a proof that our bound follows from those of [98] in the appendix.

Lemma 2.2 ([98] Theorem 2)**.** *Let $\{X_l, 1 \leq l \leq L\}$ be a completely general sequence of r.v.s. Suppose that for some fixed $\gamma > 1$, $\nu \geq \gamma$, and $C > 0$ the following condition holds:*

(i) For all $\lambda > 0$ and non-negative integers $1 \leq i \leq j \leq L$,

$$\mathbb{P}\big(|\sum_{k=i}^{j} X_k| \geq \lambda\big) \leq \big(C(j-i+1)\big)^{\gamma}\lambda^{-\nu}.$$

Then it must also hold that

$$\mathbb{P}\big(\max_{i\in[1,L]}|\sum_{k=1}^{i}X_k|\geq\lambda\big)\leq\big(6\frac{\nu+1}{\gamma-1}\big)^{\nu+1}(CL)^{\gamma}\lambda^{-\nu}.$$

2.4.2 Bound the supremum of $n't-\sum_{i=1}^{n'}N_i(t)$ over sets of consecutive integers.

In this subsection we prove a bound for the supremum term associated with $\sum_{i=1}^{n'}N_i(t)$, when evaluated at finite subsets of consecutive integer times. In particular, we will prove the following result.

Lemma 2.3. *Suppose that $\mathbb{E}[S]=1$, and that for some fixed $n'\geq 1, C_1>0, s_1>1$, and $r_1\geq s_1$, the following condition holds:*

(i) For all $t\geq 1$,

$$\mathbb{E}\big[|\sum_{i=1}^{n'}N_i(t)-n't|^{r_1}\big]\leq C_1 n'^{\frac{r_1}{2}}t^{s_1}.$$

Then it also holds that for all non-negative integers k and $\lambda>0$,

$$\mathbb{P}\big(\max_{j\in[1,k]}\big|n'j-\sum_{i=1}^{n'}N_i(j)\big|\geq\lambda\big)$$

is at most

$$\big(6\frac{r_1+1}{s_1-1}\big)^{r_1+1}C_1 n'^{\frac{r_1}{2}}k^{s_1}\lambda^{-r_1}.$$

Proof of Lemma 2.3. We proceed by verifying that for each fixed $k\geq 1$, the conditions of Lemma 2.2 hold for $\Big\{n'-\sum_{i=1}^{n'}\big(N_i(j)-N_i(j-1)\big), j=1,\ldots,k\Big\}$. Let us fix some $k\geq 1$, and non-negative integers $l\leq m\leq k$. Then for any $\lambda>0$, it follows from the fact that the given sequence of r.v.s is centered and stationary, the independence of $\{N_i(t), i\geq 1\}$, our assumptions, and Markov's inequality (after raising both sides to the r_1 power), that for all

$1 \le l \le m \le k$,

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\sum_{j=l}^{m}\Big(n' - \sum_{i=1}^{n'}\big(N_i(j) - N_i(j-1)\big)\Big)\Big| \ge \lambda\Big) \\
&\le \quad \mathbb{E}\Big[\Big|\sum_{j=l}^{m}\Big(n' - \sum_{i=1}^{n'}\big(N_i(j) - N_i(j-1)\big)\Big)\Big|^{r_1}\Big]\lambda^{-r_1} \\
&= \quad \mathbb{E}\Big[\Big|n'(m-l+1) - \sum_{i=1}^{n'}\big(N_i(m) - N_i(l-1)\big)\Big|^{r_1}\Big]\lambda^{-r_1} \\
&= \quad \mathbb{E}\Big[\Big|\sum_{i=1}^{n'} N_i(m-l+1) - n'(m-l+1)\Big|^{r_1}\Big]\lambda^{-r_1} \\
&\le \quad C_1 {n'}^{\frac{r_1}{2}}(m-l+1)^{s_1}\lambda^{-r_1}.
\end{aligned}$$

Thus we find that the conditions of Lemma 2.2 are met with $L = k$, $\{X_l, 1 \le l \le L\} = \Big\{n' - \sum_{i=1}^{n'}\big(N_i(l) - N_i(l-1)\big), l = 1, \ldots, k\Big\}$, $C = (C_1 {n'}^{\frac{r_1}{2}})^{\frac{1}{s_1}}$, $\nu = r_1$, $\gamma = s_1$, and the desired result follows. □

2.4.3 Bound the supremum of $n't - \sum_{i=1}^{n'} N_i(t)$ over intervals of length at most 1.

In this subsection we prove a bound for the supremum term associated with $\sum_{i=1}^{n'} N_i(t)$, when evaluated over intervals of length at most 1. In particular, we will prove the following result.

Lemma 2.4. *Suppose that $\mathbb{E}[S] = 1$, and that for some fixed $n' \ge 1, C_2 > 0$ and $r_2 > 2$, the following condition holds:*

(i) For all $t \in [0, 1]$,

$$\mathbb{E}\big[|\sum_{i=1}^{n'} N_i(t) - n't|^{r_2}\big] \le C_2 \max\big(n't, (n't)^{\frac{r_2}{2}}\big).$$

Then it also holds that for all $t_0 \in [0,1]$ *and* $\lambda \geq 2$*,*

$$\mathbb{P}\big(\sup_{t\in[0,t_0]} \big|n't - \sum_{i=1}^{n'} N_i(t)\big| \geq \lambda\big) \tag{2.9}$$

is at most

$$\big(24\frac{r_2+1}{r_2-2}\big)^{r_2+1} C_2 (n't_0)^{\frac{r_2}{2}} \lambda^{-r_2}.$$

Proof of Lemma 2.4. We begin by noting that it suffices to bound the supremum of interest over a suitable mesh, which follows immediately from the fact that w.p.1 $n't - \sum_{i=1}^{n'} N_i(t)$ can increase by at most 1 over any interval of length at most $(n')^{-1}$. In particular, (2.9) is at most

$$\mathbb{P}\bigg(1 + \max_{k\in[0,\lfloor n't_0\rfloor]} \big(k - \sum_{i=1}^{n'} N_i(\frac{k}{n'})\big) \geq \lambda\bigg). \tag{2.10}$$

We now verify that the conditions of Lemma 2.2 hold for $\bigg\{1-\sum_{i=1}^{n'}\big(N_i(\frac{k}{n'})-N_i(\frac{k-1}{n'})\big), k = 1,\ldots,\lfloor n't_0\rfloor\bigg\}$. Let us fix some non-negative integers $m \leq j \leq \lfloor n't_0\rfloor$ (note that if $\lfloor n't_0\rfloor < 1$ the result is trivial). Then for any $\lambda > 0$, it follows from stationary increments, centeredness, and Markov's inequality (after raising both sides to the r_2 power) that

$$\mathbb{P}\bigg(\bigg|\sum_{l=m}^{j}\bigg(1 - \sum_{i=1}^{n'}\big(N_i(\frac{l}{n'}) - N_i(\frac{l-1}{n'})\big)\bigg)\bigg| \geq \lambda\bigg) \tag{2.11}$$

is at most

$$\begin{aligned}
&\mathbb{E}\bigg[\bigg|\sum_{l=m}^{j}\bigg(1 - \sum_{i=1}^{n'}\big(N_i(\frac{l}{n'}) - N_i(\frac{l-1}{n'})\big)\bigg)\bigg|^{r_2}\bigg]\lambda^{-r_2} \\
&= \quad \mathbb{E}\bigg[\big|(j-m+1) - \sum_{i=1}^{n'} N_i(\frac{j-m+1}{n'})\big|^{r_2}\bigg]\lambda^{-r_2},
\end{aligned}$$

which by our assumptions (and noting that in this case the $n't$ appearing in our assumptions equals $j-m+1$) is at most $C_2 \max\big(j-m+1, (j-m+1)^{\frac{r_2}{2}}\big)$. Since $j-m+1$ is a non-negative integer and $\frac{r_2}{2} \geq 1$, it follows that (2.11) is at most $C_2(j-m+1)^{\frac{r_2}{2}}\lambda^{-r_2}$. We

thus find that the conditions of Lemma 2.2 are met with $L = \lfloor n't_0 \rfloor$, $\{X_l, 1 \leq l \leq L\} = \left\{1 - \sum_{i=1}^{n'} \left(N_i(\frac{l}{n'}) - N_i(\frac{l-1}{n'})\right), l = 1, \ldots, \lfloor n't_0 \rfloor\right\}$, $C = (C_2)^{\frac{2}{r_2}}$, $\nu = r_2$, $\gamma = \frac{r_2}{2}$. Thus for all $z > 0$,

$$\mathbb{P}\Big(\max_{k \in [0, \lfloor n't_0 \rfloor]} \big(k - \sum_{i=1}^{n'} N_i(\frac{k}{n'})\big) \geq z\Big) \leq \big(6\frac{r_2+1}{\frac{r_2}{2}-1}\big)^{r_2+1} C_2 (n't_0)^{\frac{r_2}{2}} z^{-r_2}.$$

It then follows from (2.10), and the fact that $\lambda \geq 2$ implies $(\lambda - 1)^{-r_2} \leq 2^{r_2} \lambda^{-r_2}$, that (2.9) is at most

$$\begin{aligned} & 2^{r_2} \times \big(6\frac{r_2+1}{\frac{r_2}{2}-1}\big)^{r_2+1} \times C_2 \times (n't_0)^{\frac{r_2}{2}} \lambda^{-r_2} \\ \leq \quad & \big(24\frac{r_2+1}{r_2-2}\big)^{r_2+1} C_2 (n't_0)^{\frac{r_2}{2}} \lambda^{-r_2}, \end{aligned}$$

completing the proof. □

2.4.4 Bound $\sup_{t \geq 0} \left(n't - \sum_{i=1}^{n'} N_i(t) - \frac{1}{2}(n' - \mu_{A,n'})t\right)$ by combining Lemmas 2.3 and 2.4.

We now combine our two bounds for the supremum associated with $n't - \sum_{i=1}^{n'} N_i(t)$, namely Lemma 2.3 which provides bounds over sets of integer times, and Lemma 2.4 which provides bounds over intervals of length at most 1. We proceed by proving a very general result for the all-time supremum of continuous-time random walks with stationary increments and negative drift, which exactly converts appropriate bounds for the supremum over consecutive integers and over intervals of length at most 1 to bounds for the all-time supremum. We note that similar arguments have been used to bound all-time suprema of stochastic processes (cf. [100]), also in the heavy-tailed setting (cf. [101]). To allow for minimal assumptions, and to allow for the fully range of applicability to our setting of interest (and for completeness), we include a self-contained exposition and proof. We will rely on the following result, whose proof we defer to the appendix.

Lemma 2.5. *Let $\{\phi(t), t \geq 0\}$ be a stochastic process with stationary increments such that $\phi(0) = 0$. Here, stationary increments means that for all $s_0 \geq 0$, $\{\phi(s+s_0)-\phi(s_0), s \geq 0\}$ has the same distribution (on the process level) as $\{\phi(s), s \geq 0\}$. Suppose there exist strictly positive finite constants H_1, H_2, s, r_1, r_2 and $Z \geq 0$ such that $r_1 > s > 1$ and $r_2 > 2$, and the following two conditions hold:*

(i) For all integers $m \geq 1$ and real numbers $\lambda \geq Z$,

$$\mathbb{P}(\max_{j\in\{0,\ldots,m\}} \phi(j) \geq \lambda) \leq H_1 m^s \lambda^{-r_1}.$$

(ii) For all $t_0 \in (0, 1]$ and $\lambda \geq Z$,

$$\mathbb{P}(\sup_{0\leq t\leq t_0} \phi(t) \geq \lambda) \leq H_2 t_0^{\frac{r_2}{2}} \lambda^{-r_2}.$$

Then for any drift parameter $\nu > 0$, and all $\lambda \geq 4Z$, $\mathbb{P}\left(\sup_{t\geq 0}(\phi(t) - \nu t) \geq \lambda\right)$ is at most

$$(1 + \frac{1}{r_1 - s})4^{r_1+r_2+2}\left(H_1\nu^{-s}\lambda^{-(r_1-s)} + H_2(\lambda\nu)^{-\frac{r_2}{2}}\right)$$

With Lemma 2.5 in hand, we now combine with Lemmas 2.3 and 2.4 to prove the following bound for $\sup_{t\geq 0}\left(n't - \sum_{i=1}^{n'} N_i(t) - \nu t\right)$. We note that ultimately we will take $\nu = \frac{1}{2}(n' - \mu_A)$, but here we prove the result for general drift.

Lemma 2.6. *Suppose that $\mathbb{E}[S] = 1$, and that for some fixed $n' \geq 1, C_1, C_2 > 0; r_1 > s_1 > 1$; and $r_2 > 2$:*

(i) For all $t \geq 1$,

$$\mathbb{E}\big[|\sum_{i=1}^{n'} N_i(t) - n't|^{r_1}\big] \leq C_1 n'^{\frac{r_1}{2}} t^{s_1}.$$

(ii) For all $t \in [0,1]$*,*

$$\mathbb{E}\big[|\sum_{i=1}^{n'} N_i(t) - n't|^{r_2}\big] \leq C_2 \max\big(n't, (n't)^{\frac{r_2}{2}}\big).$$

Then for all $\nu > 0$ *and* $\lambda \geq 8$*,*

$$\mathbb{P}\left(\sup_{t\geq 0}\left(n't - \sum_{i=1}^{n'} N_i(t) - \nu t\right) \geq \lambda\right)$$

is at most

$$\left(\frac{100(r_1+r_2)^3}{(s_1-1)(r_1-s_1)(r_2-2)}\right)^{r_1+r_2+2}\left(C_1 n'^{\frac{r_1}{2}}\nu^{-s_1}\lambda^{-(r_1-s_1)} + C_2 n'^{\frac{r_2}{2}}(\lambda\nu)^{-\frac{r_2}{2}}\right).$$

Proof of Lemma 2.6. By our assumptions and Lemma 2.3, for all non-negative integers k and $\lambda > 0$,

$$\mathbb{P}\big(\max_{j\in[1,k]}\big|n'\mu_S j - \sum_{i=1}^{n'} N_i(j)\big| \geq \lambda\big) \leq \big(6\frac{r_1+1}{s_1-1}\big)^{r_1+1} C_1 n'^{\frac{r_1}{2}} k^{s_1}\lambda^{-r_1}.$$

Next, by our assumptions and Lemma 2.4, for all $t_0 \in [0,1]$ and $\lambda \geq 2$,

$$\mathbb{P}\big(\sup_{t\in[0,t_0]}\big|n'\mu_S t - \sum_{i=1}^{n'} N_i(t)\big| \geq \lambda\big) \leq \big(24\frac{r_2+1}{r_2-2}\big)^{r_2+1} C_2 n'^{\frac{r_2}{2}} t_0^{\frac{r_2}{2}}\lambda^{-r_2}.$$

It then follows from our assumptions that the conditions of Lemma 2.5 are met with $\phi(t) = n't - \sum_{i=1}^{n'} N_i(t)$, $s = s_1, r_1, r_2, \nu$ their given values, $Z = 2$,

$$H_1 = \big(6\frac{r_1+1}{s_1-1}\big)^{r_1+1} C_1 n'^{\frac{r_1}{2}} \quad , \quad H_2 = \big(24\frac{r_2+1}{r_2-2}\big)^{r_2+1} C_2 n'^{\frac{r_2}{2}}.$$

Combining the above with the implications of Lemma 2.5 and some straightforward algebra completes the proof. □

2.4.5 Bound the supremum of $k - \mu_{A,n'} \sum_{i=1}^{k} A_{i,n'}$ over sets of consecutive integers

In this subsection we prove a bound for the supremum of $k - \mu_{A,n'} \sum_{i=1}^{k} A_{i,n'}$, as an intermediate step towards bounding the supremum of $A_{o,n'}(t) - \mu_{A,n'}t - \frac{1}{2}(n' - \mu_{A,n'})t$. In particular, we will prove the following result.

Lemma 2.7. *Suppose that $0 < \mu_A < \infty$, and for some fixed $C_3 > 0, s_3 > 1$, and $r_3 \geq s_3$, the following condition holds:*

(i) For all $k \geq 1$,

$$\mathbb{E}\big[|k - \sum_{i=1}^{k}(\mu_A A_i)|^{r_3}\big] \leq C_3 k^{s_3}.$$

Then for all $n' \geq 1$ and non-negative integers k and $\lambda > 0$,

$$\mathbb{P}\big(\max_{j\in[1,k]} \big|j - \sum_{i=1}^{j}(\mu_{A,n'} A_{i,n'})\big| \geq \lambda\big)$$

is at most

$$\big(6\frac{r_3+1}{s_3-1}\big)^{r_3+1} C_3 k^{s_3} \lambda^{-r_3}.$$

Proof of Lemma 2.7. We proceed by verifying that for each fixed $k \geq 1$, the conditions of Lemma 2.2 hold for $\Big\{1 - \mu_{A,n'} A_{j,n'}, j = 1, \ldots, k\Big\}$. First, note that

$$\mathbb{E}\big[|k - \sum_{i=1}^{k}(\mu_{A,n'} A_{i,n'})|^{r_3}\big] = \mathbb{E}\big[|k - \sum_{i=1}^{k}(\mu_A A_i)|^{r_3}\big]. \tag{2.12}$$

Let us fix some $k \geq 1$, and non-negative integers $l \leq m \leq k$. Then for any $\lambda > 0$, it follows from the fact that the given sequence of r.v.s is centered and stationary, the independence of $\{A_{i,n'}, i \geq 1\}$, (2.12), and Markov's inequality (after raising both sides to the r_3 power),

that

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j=l}^{m}\left(1-\mu_{A,n'}A_{j,n'}\right)\right|\geq\lambda\right)\\
&\leq\quad\mathbb{E}\left[\left|\sum_{j=l}^{m}\left(1-\mu_{A,n'}A_{j,n'}\right)\right|^{r_3}\right]\lambda^{-r_3}\\
&=\quad\mathbb{E}\left[\left|\sum_{i=1}^{m-l+1}\left(\mu_{A,n'}A_{i,n'}\right)-(m-l+1)\right|^{r_3}\right]\lambda^{-r_3}\\
&\leq\quad C_3(m-l+1)^{s_3}\lambda^{-r_3}.
\end{aligned}$$

Thus we find that the conditions of Lemma 2.2 are met with $L=k$, $\{X_l, 1\leq l\leq L\}=\left\{1-\mu_{A,n'}A_{l,n'}, l=1,\ldots,k\right\}$, $C=(C_3)^{\frac{1}{s}}$, $\nu=r_3$, $\gamma=s_3$, and the desired result follows. □

2.4.6 Bound $\sup_{t\geq 0}\left(A_{o,n'}(t)-\mu_{A,n'}t-\frac{1}{2}(n'-\mu_{A,n'})t\right)$ using Lemma 2.7.

We now use Lemma 2.7 to bound $\sup_{t\geq 0}\left(A_{o,n'}(t)-\mu_{A,n'}t-\frac{1}{2}(n'-\mu_{A,n'})t\right)$. Here we prove the result for general linear drift, but will later connect back to the desired drift $\frac{1}{2}(n'-\mu_{A,n'})$. We proceed in three steps. First, we relate the desired supremum to a discrete-time supremum associated with $k-\mu_A\sum_{i=1}^{k}A_i$. In particular, we begin with the following lemma.

Lemma 2.8. *Suppose that* $0<\mu_A<\infty$. *Then for all* $n'\geq 1, \nu>0$ *and* $\lambda>0$,

$$\mathbb{P}\left(\sup_{t\geq 0}\left(A_{o,n'}(t)-\mu_{A,n'}t-\nu t\right)\geq\lambda\right) \tag{2.13}$$

equals

$$\mathbb{P}\left(\sup_{k\geq 0}\left(k-\mu_{A,n'}\sum_{i=1}^{k}A_{i,n'}-\frac{\nu}{\mu_{A,n'}+\nu}k\right)\geq\lambda(1+\frac{\nu}{\mu_{A,n'}})^{-1}\right). \tag{2.14}$$

Proof of Lemma 2.8. As $\{A_{o,n'}(t)-\mu_{A,n'}t-\nu t, t\geq 0\}$ jumps up only at times $\{\sum_{i=1}^{k}A_{i,n'}, k\geq$

$1\}$ and at all other times drifts downward at linear rate $-(\mu_{A,n'}+\nu)$, we conclude that we may examine the relevant supremum only at times $\{\sum_{i=1}^{k} A_{i,n'}, k \geq 0\}$, from which it follows that (2.13) equals

$$\mathbb{P}\bigg(\sup_{k\geq 0} \big(k - (\mu_{A,n'}+\nu)\sum_{i=1}^{k} A_{i,n'}\big) \geq \lambda \bigg). \tag{2.15}$$

Further observing that

$$\begin{aligned} k - (\mu_{A,n'}+\nu)\sum_{i=1}^{k} A_{i,n'} &= (1+\frac{\nu}{\mu_{A,n'}})k - (\mu_{A,n'}+\nu)\sum_{i=1}^{k} A_{i,n'} - \frac{\nu}{\mu_{A,n'}}k \\ &= (1+\frac{\nu}{\mu_{A,n'}})\big(k - \mu_{A,n'}\sum_{i=1}^{k} A_{i,n'} - \frac{\nu}{\mu_{A,n'}+\nu}k\big) \end{aligned}$$

completes the proof. □

Second, we prove a general result for the all-time supremum of discrete-time random walks with stationary increments and negative drift, in analogy with Lemma 2.5, which we will use to analyze (2.14). In particular, we prove the following result, whose proof we defer to the appendix.

Lemma 2.9. *Let $\{\phi(k), k \geq 0\}$ be a discrete-time stochastic process with stationary increments such that $\phi(0) = 0$. Here, stationary increments means that for all integers $k_0 \geq 0$, $\{\phi(k+k_0) - \phi(k_0), k \geq 0\}$ has the same distribution (on the process level) as $\{\phi(k), k \geq 0\}$. Suppose there exist strictly positive finite constants H_3, s_3, r_3 such that $r_3 > s_3 \geq 1$, and the following condition holds:*

(i) For all integers $m \geq 1$ and $\lambda > 0$,

$$\mathbb{P}(\max_{j\in\{0,\ldots,m\}} \phi(j) \geq \lambda) \leq H_3 m^{s_3} \lambda^{-r_3}.$$

Then for any drift parameter $\nu > 0$, and all $\lambda > 0$, $\mathbb{P}\Big(\sup_{k\geq 0}(\phi(k) - \nu k) \geq \lambda \Big)$ is at most

$$16H_3 4^{r_3}(1 + \frac{1}{r_3 - s_3})\nu^{-s_3}\lambda^{-(r_3-s_3)}.$$

Finally, we combine Lemmas 2.8 and 2.9 to bound $\sup_{t\geq 0}\Big(A_{o,n'}(t) - \mu_{A,n'}t - \frac{1}{2}(n' - \mu_{A,n'})t\Big)$, proving the following.

Lemma 2.10. *Suppose that $0 < \mu_A < \infty$, and that for some fixed $C_3 > 0$ and $r_3 > s_3 > 1$, the following condition holds:*

(i) For all $k \geq 1$,

$$\mathbb{E}\big[|k - \mu_A \sum_{i=1}^{k} A_i|^{r_3}\big] \leq C_3 k^{s_3}.$$

Then for all $\nu > 0$ and $\lambda > 0$,

$$\mathbb{P}\Big(\sup_{t\geq 0} \big(A_{o,n'}(t) - \mu_{A,n'}t - \nu t\big) \geq \lambda \Big)$$

is at most

$$\Big(\frac{10^3(r_3+1)^2}{(s_3-1)(r_3-s_3)}\Big)^{r_3+1} C_3\nu^{-s_3}(\mu_{A,n'} + \nu)^{r_3}\mu_{A,n'}^{-(r_3-s_3)}\lambda^{-(r_3-s_3)}.$$

Proof of Lemma 2.10. By Lemma 2.8, it suffices to bound

$$\mathbb{P}\Big(\sup_{k\geq 0} \Big(k - \mu_{A,n'}\sum_{i=1}^{k} A_{i,n'} - \frac{\nu}{\mu_{A,n'} + \nu}k\Big) \geq \lambda(1 + \frac{\nu}{\mu_{A,n'}})^{-1} \Big) \tag{2.16}$$

Our assumptions, (2.12), and Lemma 2.7 ensure that for all non-negative integers k and $\lambda' > 0$,

$$\mathbb{P}\big(\max_{j\in[1,k]} \big|j - \sum_{i=1}^{j}(\mu_{A,n'}A_{i,n'})\big| \geq \lambda' \big) \tag{2.17}$$

is at most

$$\big(6\frac{r_3+1}{s_3-1}\big)^{r_3+1}C_3k^{s_3}(\lambda')^{-r_3}.$$

It then follows from our assumptions that the conditions of Lemma 2.9 are met with $\phi(k) = k - \mu_{A,n'}\sum_{i=1}^{k}A_{i,n'}$, s_3, r_3 there given values,

$$H_3 = \big(6\frac{r_3+1}{s_3-1}\big)^{r_3+1}C_3.$$

We conclude (after setting the drift parameter equal to $\frac{\nu}{\mu_{A,n'}+\nu}$ and the target level equal to $\lambda(1+\frac{\nu}{\mu_{A,n'}})^{-1}$), that for all $\lambda > 0$, (2.16) is at most

$$16\big(6\frac{r_3+1}{s_3-1}\big)^{r_3+1}C_3 4^{r_3}(1+\frac{1}{r_3-s_3})(\frac{\nu}{\mu_{A,n'}+\nu})^{-s_3}\big(\lambda(1+\frac{\nu}{\mu_{A,n'}})^{-1}\big)^{-(r_3-s_3)}.$$

Combining with some straightforward algebra completes the proof. □

2.4.7 Proof of Theorem 2.4.

With Lemmas 2.6 and 2.10 in hand, we now complete the proof of Theorem 2.4.

Proof of Theorem 2.4. Using Lemma 2.6 to bound (2.8), and Lemma 2.10 to bound (2.7), combined with Lemma 2.1 and some straightforward algebra, we conclude that for all $x \geq 16$, $\mathbb{P}\bigg(\sup_{t\geq 0}\Big(A(t) - \sum_{i=1}^{n'}N_i(t)\Big) \geq x\bigg)$ is at most

$$\begin{aligned}&\bigg(\frac{100(r_1+r_2)^3}{(s_1-1)(r_1-s_1)(r_2-2)}\bigg)^{r_1+r_2+2} \\ &\times\quad \bigg(C_1{n'}^{\frac{r_1}{2}}\big(\frac{1}{2}(n'-\mu_{A,n'})\big)^{-s_1}(\frac{x}{2})^{-(r_1-s_1)} + C_2{n'}^{\frac{r_2}{2}}\big(\frac{x}{2}\times\frac{1}{2}(n'-\mu_{A,n'})\big)^{-\frac{r_2}{2}}\bigg)\end{aligned} \tag{2.18}$$

+

$$\left(\frac{10^3(r_3+1)^2}{(s_3-1)(r_3-s_3)}\right)^{r_3+1} \tag{2.19}$$
$$\times \quad C_3\big(\frac{1}{2}(n'-\mu_{A,n'})\big)^{-s_3}\big(\frac{1}{2}(n'+\mu_{A,n'})\big)^{r_3}\mu_{A,n'}^{-(r_3-s_3)}(\frac{x}{2}-1)^{-(r_3-s_3)}.$$

It follows from some straightforward algebra, and the assumption that $n' > \mu_{A,n'}$, that (2.18) is at most

$$\left(\frac{16\times 10^3(r_1+r_2+r_3)^5}{(s_1-1)(s_3-1)(r_1-s_1)(r_3-s_3)(r_2-2)}\right)^{r_1+r_2+r_3+1}(1+C_1)(1+C_2)(1+C_3)$$
$$\times \quad \left(n'^{\frac{r_1}{2}}(n'-\mu_{A,n'})^{-s_1}x^{-(r_1-s_1)}+n'^{\frac{r_2}{2}}(n'-\mu_{A,n'})^{-\frac{r_2}{2}}x^{-\frac{r_2}{2}}\right),$$

and (2.19) is at most

$$\left(\frac{64\times 10^3(r_1+r_2+r_3)^5}{(s_1-1)(s_3-1)(r_1-s_1)(r_3-s_3)(r_2-2)}\right)^{r_1+r_2+r_3+1}(1+C_1)(1+C_2)(1+C_3)$$
$$\times \quad (n'-\mu_{A,n'})^{-s_3}(n')^{r_3}\mu_{A,n'}^{-(r_3-s_3)}x^{-(r_3-s_3)}.$$

Combining the above with some straightforward algebra completes the proof. □

2.5 Making Theorem 2.4 completely explicit (proof of Theorem 2.1).

In this section we show that the relevant (pooled) renewal processes satisfy the conditions of Theorem 2.4 for certain explicit constants (assuming finite second moment), and use the corresponding explicit result of Theorem 2.4, combined with the stochastic comparison results Theorems 2.2 and 2.3, to complete the proof of Theorem 2.1.

2.5.1 Bounding the central moments of $\sum_{i=1}^k N_i(t)$ for $t \geq 1$

In this subsection we bound the central moments of $\sum_{i=1}^k N_i(t)$ for $t \geq 1$. In particular, we will prove the following.

Lemma 2.11. *Suppose that* $\mathbb{E}[S] = 1$*, and that* $\mathbb{E}[S^r] < \infty$ *for some* $r \geq 2$*. Then for all* $k \geq 1$*,* $t \geq 1$*, and* $\theta > 0$*,*

$$\mathbb{E}\big[\big|\sum_{i=1}^{k} N_i(t) - kt\big|^r\big] \leq \mathbb{E}[S^r]\exp(\theta)\big(\frac{10^8 r^3}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+2} k^{\frac{r}{2}} t^{\frac{r}{2}}.$$

Our proof of Lemma 2.11 proceeds in several steps. First, we bound the rth central moment of $N_o(t)$, showing that this moment scales (with t) like $t^{\frac{r}{2}}$ and providing a completely explicit bound along these lines. Our proof can essentially be viewed as "making completely explicit", e.g. all constants explicitly worked out, the approach to bounding the central moments of a renewal process sketched in [102]. As noted in [102] (and used in [63]), a non-explicit bound proving that the rth central moment indeed scales asymptotically (with t) like $t^{\frac{r}{2}}$ was first proven in [103]. To our knowledge such a completely explicit bound is new, and may prove useful in other settings. In particular, we begin by proving the following.

Lemma 2.12. *Suppose that* $\mathbb{E}[S] = 1$*, and that* $\mathbb{E}[S^r] < \infty$ *for some* $r \geq 2$*. Then for all* $t \geq 1$ *and* $\theta > 0$*,*

$$\mathbb{E}\left[\left|N_o(t) - t\right|^r\right] \leq \exp(\theta)\mathbb{E}[S^r]\big(\frac{10^5 r^2}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+1} t^{\frac{r}{2}}.$$

We begin with some preliminary technical results. First, we recall the the celebrated Burkholder-Rosenthal Inequality for bounding the moments of a martingale. We state a particular variant (chosen largely for simplicity, although tighter bounds are known) given in [104].

Lemma 2.13 (Burkholder-Rosenthal Inequality, [104])**.** *Let* $\{X_i, i \geq 1\}$ *be a martingale difference sequence w.r.t. the filtration* $\{\mathcal{F}_i, i \geq 0\}$*. Namely, we have that* $\{X_i, i \geq 1\}$ *is adapted to* $\{\mathcal{F}_i, i \geq 0\}$*;* $\mathbb{E}[|X_i|] < \infty$ *for all* $i \geq 1$*; and* $\mathbb{E}[X_i|\mathcal{F}_{i-1}] = 0$ *for all* $i \geq 1$*. Then*

for all $r \geq 2$*,*

$$\left(\mathbb{E}\big[\big|\sum_{i=1}^{\infty} X_i\big|^r\big]\right)^{\frac{1}{r}}$$

is at most

$$10r\left(\mathbb{E}\left[\left(\sum_{i=1}^{\infty}\mathbb{E}[X_i^2|\mathcal{F}_{i-1}]\right)^{\frac{r}{2}}\right]\right)^{\frac{1}{r}} + 10r\left(\mathbb{E}\big[\sup_{i\geq 1}|X_i^r|\big]\right)^{\frac{1}{r}}.$$

Since for any sequence of r.v.s $\{Z_i, i = 1, \ldots, n\}$ and $r \geq 1$ it follows from convexity that w.p.1

$$\big|\sum_{i=1}^{n} Z_i\big|^r \leq n^{r-1}\sum_{i=1}^{n}|Z_i|^r, \tag{2.20}$$

we deduce the following corollary.

Corollary 2.6. *Under the same definitions and assumptions as Lemma 2.13, for all* $r \geq 2$*,* $\mathbb{E}\big[\big|\sum_{i=1}^{\infty} X_i\big|^r\big]$ *is at most*

$$(20r)^r\mathbb{E}\left[\left(\sum_{i=1}^{\infty}\mathbb{E}[X_i^2|\mathcal{F}_{i-1}]\right)^{\frac{r}{2}}\right] + (20r)^r\mathbb{E}\big[\sup_{i\geq 1}|X_i^r|\big].$$

We next recall a certain inequality for the non-central moments of $N_o(t)$, proven in [102] Equation 5.11.

Lemma 2.14 ([102] Equation 5.11)**.** *Suppose that* $0 < \mu_S < \infty$*. Then for all* $p > 0$ *and* $t \geq 1$*,*

$$\mathbb{E}\Big[\big(N_o(t)+1\big)^p\Big] \leq (2t)^p\mathbb{E}\Big[\big(N_o(1)+1\big)^p\Big].$$

We also prove the following bounds for moments of $N_o(1)$, whose proof we defer to the appendix.

Lemma 2.15. *Suppose that* $0 < \mu_S < \infty$*. Then for all* $p \geq 1$ *and* $\theta > 0$*,*

$$\mathbb{E}\big[\big(N_o(1)\big)^p\big] \leq \exp(\theta)\big(\frac{24p}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{p+2}.$$

Combining Lemmas 2.14 and 2.15 with (2.20) and some straightforward algebra, we

come to the following corollary.

Corollary 2.7. *Suppose that* $0 < \mu_S < \infty$. *Then for all* $p \geq 1$, $t \geq 1$, *and* $\theta > 0$,

$$\mathbb{E}\Big[\big(N_o(t)+1\big)^p\Big] \leq \exp(\theta)\big(\frac{100p}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{p+2}t^p.$$

With Lemmas 2.13 - 2.15 and Corollary 2.7 in hand, we now complete the proof of Lemma 2.12.

Proof of Lemma 2.12. By definition (as is well-known), $N_o(t)+1 = \min\{n \geq 1 : \sum_{i=1}^n S_i > t\}$ is a stopping time w.r.t. the natural filtration generated by $\{S_i, i \geq 1\}$. By the triangle inequality, w.p.1

$$\begin{aligned}
\big|N_o(t)-t\big| &= \big|\big(N_o(t)+1\big)-t-1\big| \\
&\leq \big|\big(N_o(t)+1\big)-t\big|+1 \\
&\leq \Big|\sum_{i=1}^{N_o(t)+1} S_i - \big(N_o(t)+1\big)\Big| + \Big|\sum_{i=1}^{N_o(t)+1} S_i - t\Big| + 1. \qquad (2.21)
\end{aligned}$$

It then follows from (2.20) and (2.21) that $\mathbb{E}\Big[\Big|N_o(t)-t\Big|^r\Big]$ is at most

$$\begin{aligned}
& 3^{r-1}\mathbb{E}\Big[\Big|\sum_{i=1}^{N_o(t)+1} S_i - \big(N_o(t)+1\big)\Big|^r\Big] && (2.22)\\
+\ & 3^{r-1}\mathbb{E}\Big[\Big|\sum_{i=1}^{N_o(t)+1} S_i - t\Big|^r\Big] && (2.23)\\
+\ & 3^{r-1}. && (2.24)
\end{aligned}$$

We next bound

$$\mathbb{E}\Big[\Big|\sum_{i=1}^{N_o(t)+1} S_i - \big(N_o(t)+1\big)\Big|^r\Big], \qquad (2.25)$$

and proceed by applying the celebrated Burkholder-Rosenthal Inequality. In particular, we will use Corollary 2.6 to bound (2.25). First, we rewrite (2.25) in terms of an appropriate

martingale difference sequence. Namely, note that (2.25) equals

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i=1}^{\infty} S_i I\big(N_o(t)+1\geq i\big) - \sum_{i=1}^{\infty} I\big(N_o(t)+1\geq i\big)\right|^r\right] \\
&= \mathbb{E}\left[\left|\sum_{i=1}^{\infty} (S_i-1) I\big(N_o(t)+1\geq i\big)\right|^r\right]. \qquad (2.26)
\end{aligned}$$

We now prove that $\{(S_i-1)I\big(N_o(t)+1\geq i\big), i\geq 1\}$ is a martingale difference sequence w.r.t. the filtration $\{\sigma(S_1,\ldots,S_i), i\geq 1\}$. Finite expecations and measurability are trivial. Furthermore, since $I\big(N_o(t)+1\geq i\big)$ is $\sigma(S_1,\ldots,S_{i-1})$-measurable (due to the greater than or equal to sign), it follows from independence and the basic properties of conditional expectation that w.p.1

$$\begin{aligned}
&\mathbb{E}\Big[(S_i-1)I\big(N_o(t)+1\geq i\big)|\sigma(S_1,\ldots,S_{i-1})\Big] \\
&= I\big(N_o(t)+1\geq i\big)\mathbb{E}\Big[(S_i-1)|\sigma(S_1,\ldots,S_{i-1})\Big] \\
&= I\big(N_o(t)+1\geq i\big)\mathbb{E}\big[S_i-1\big] \quad = \quad 0.
\end{aligned}$$

Thus we find that the conditions of Corollary 2.6 are satisfied with $X_i = (S_i-1)I\big(N_o(t)+1\geq i\big), \mathcal{F}_i = \sigma(S_1,\ldots,S_i)$. Before stating the given implication, we first show that several

resulting terms can be simplified. First, note that

$$\begin{aligned}
&\mathbb{E}\Bigg[\Bigg(\sum_{i=1}^{\infty}\mathbb{E}\Bigg[\Big((S_i-1)I\big(N_o(t)+1\geq i\big)\Big)^2\Bigg|\sigma(S_1,\ldots,S_{i-1})\Bigg]\Bigg)^{\frac{r}{2}}\Bigg]\\
&= \mathbb{E}\Bigg[\Bigg(\sum_{i=1}^{\infty}\mathbb{E}\Bigg[(S_i-1)^2I\big(N_o(t)+1\geq i\big)\Bigg|\sigma(S_1,\ldots,S_{i-1})\Bigg]\Bigg)^{\frac{r}{2}}\Bigg]\\
&= \mathbb{E}\Bigg[\Bigg(\sum_{i=1}^{\infty}I\big(N_o(t)+1\geq i\big)\mathbb{E}\Bigg[(S_i-1)^2\Bigg|\sigma(S_1,\ldots,S_{i-1})\Bigg]\Bigg)^{\frac{r}{2}}\Bigg]\\
&= \mathbb{E}\Bigg[\Bigg(\sum_{i=1}^{\infty}I\big(N_o(t)+1\geq i\big)\mathbb{E}\big[(S-1)^2\big]\Bigg)^{\frac{r}{2}}\Bigg]\\
&= \Big(\mathbb{E}\big[(S-1)^2\big]\Big)^{\frac{r}{2}}\mathbb{E}\Bigg[\Bigg(\sum_{i=1}^{\infty}I\big(N_o(t)+1\geq i\big)\Bigg)^{\frac{r}{2}}\Bigg]\\
&= \Big(\mathbb{E}\big[(S-1)^2\big]\Big)^{\frac{r}{2}}\mathbb{E}\Big[\big(N_o(t)+1\big)^{\frac{r}{2}}\Big]\\
&\leq \mathbb{E}\big[|S-1|^r\big]\mathbb{E}\Big[\big(N_o(t)+1\big)^{\frac{r}{2}}\Big], \qquad (2.27)
\end{aligned}$$

the final inequality following from Jensen's inequality (applicable since $r \geq 2$). Second, note that

$$\begin{aligned}
&\mathbb{E}\Bigg[\sup_{i\geq 1}\Bigg|\Big((S_i-1)I\big(N_o(t)+1\geq i\big)\Big)^r\Bigg|\Bigg]\\
&\leq \mathbb{E}\Bigg[\sum_{i=1}^{\infty}I\big(N_o(t)+1\geq i\big)\big|S_i-1\big|^r\Bigg]\\
&= \mathbb{E}\Bigg[\sum_{i=1}^{N_o(t)+1}\big|S_i-1\big|^r\Bigg]\\
&= \mathbb{E}\big[N_o(t)+1\big]E\big[\big|S-1\big|^r\big], \qquad (2.28)
\end{aligned}$$

the final inequality following from Wald's identity. Combining (2.27) and (2.28) with the fact that the conditions of Corollary 2.6 are satisfied with $X_i = (S_i-1)I\big(N_o(t)+1 \geq i\big), \mathcal{F}_i = \sigma(S_1,\ldots,S_i)$, and the fact that $\mathbb{E}\big[N_o(t)+1\big] \leq \mathbb{E}\Big[\big(N_o(t)+1\big)^{\frac{r}{2}}\Big]$ (since $r \geq 2$),

we conclude that (2.25) is at most

$$2(20r)^r \mathbb{E}\big[|S-1|^r\big]\mathbb{E}\Big[\big(N_o(t)+1\big)^{\frac{r}{2}}\Big]. \tag{2.29}$$

Combining (2.29) and Corollary 2.7, we conclude that (2.25) is at most

$$2(20r)^r \mathbb{E}\big[|S-1|^r\big]\exp(\theta)\bigg(\frac{50r}{1-\mathbb{E}[\exp(-\theta S)]}\bigg)^{\frac{r}{2}+2} t^{\frac{r}{2}} \tag{2.30}$$

$$\leq \quad \exp(\theta)\bigg(\frac{2\times 10^3 r^2}{1-\mathbb{E}[\exp(-\theta S)]}\bigg)^{r+1} \mathbb{E}\big[|S-1|^r\big] t^{\frac{r}{2}}. \tag{2.31}$$

We next bound (2.23), by bounding

$$\mathbb{E}\bigg[\Big|\sum_{i=1}^{N_o(t)+1} S_i - t\Big|^r\bigg]. \tag{2.32}$$

By definition, $\sum_{i=1}^{N_o(t)+1} S_i - t$ is the residual life of the renewal process N_o at time t, i.e. the remaining time until the next renewal (at time t), and it follows that w.p.1

$$\begin{aligned}\Big|\sum_{i=1}^{N_o(t)+1} S_i - t\Big|^r &\leq S^r_{N_o(t)+1} \\ &\leq \sum_{i=1}^{N_o(t)+1} S_i^r.\end{aligned}$$

Combining with Wald's identity, we conclude that (2.32) is at most $\mathbb{E}\big[N_o(t)+1\big]\mathbb{E}[S^r]$, which by Corollary 2.7 is at most

$$\exp(\theta)\big(\frac{100}{1-\mathbb{E}\big[\exp(-\theta S)\big]}\big)^3 t\mathbb{E}[S^r]. \tag{2.33}$$

Using (2.31) to bound (2.25) (providing a bound for (2.22)), and (2.33) to bound (2.32)

(providing a bound for (2.23)), and applying (2.20) and some straightforward algebra, we find that $\mathbb{E}\left[\left|N_o(t)-t\right|^r\right]$ is at most

$$\begin{aligned}
& 3^{r-1}\exp(\theta)\big(\frac{2\times 10^3 r^2}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+1}\mathbb{E}\big[|S-1|^r\big]t^{\frac{r}{2}} \\
& \quad + \quad 3^{r-1}\exp(\theta)\big(\frac{100}{1-\mathbb{E}\big[\exp(-\theta S)\big]}\big)^{3}t\mathbb{E}[S^r] \quad + \quad 3^{r-1} \\
\leq \quad & 3^{r-1}\exp(\theta)\big(\frac{2\times 10^3 r^2}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+1}t^{\frac{r}{2}}\bigg(\mathbb{E}\big[|S-1|^r\big]+\mathbb{E}[S^r]\bigg)+3^{r-1} \\
\leq \quad & 3^{r-1}\exp(\theta)\big(\frac{2\times 10^3 r^2}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+1}t^{\frac{r}{2}}\bigg(2^{r-1}\big(\mathbb{E}[S^r]+1\big)+\mathbb{E}[S^r]\bigg)+3^{r-1} \\
\leq \quad & 3^{r-1}\exp(\theta)\big(\frac{2\times 10^3 r^2}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+1}t^{\frac{r}{2}}2^{r+1}\mathbb{E}[S^r]+3^{r-1} \\
\leq \quad & \exp(\theta)\mathbb{E}[S^r]\big(\frac{10^5 r^2}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+1}t^{\frac{r}{2}}.
\end{aligned}$$

Combining the above completes the proof. □

We now extend Lemma 2.12 to the corresponding equilibrium renewal process. We note that given the results of Lemma 2.12, such an extension follows nearly identically to the proof of Lemma 8 of [63]. However, as we wish to make all quantities compeltely explicit, we include a self-contained proof in the appendix.

Corollary 2.8. *Suppose that* $\mathbb{E}[S]=1$*, and that* $\mathbb{E}[S^r]<\infty$ *for some* $r\geq 2$*. Then for all* $t\geq 1$ *and* $\theta>0$*,*

$$\mathbb{E}\left[\left|N_1(t)-t\right|^r\right]\leq \mathbb{E}[S^r]\exp(\theta)\big(\frac{10^7 r^2}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+2}t^{\frac{r}{2}}.$$

Before completing the proof of Lemma 2.11, we recall the celebrated Marcinkiewicz-Zygmund inequality, a close relative of the Rosenthal inequality. The precise result which we will use follows immediately from [105] Theorem 2, and we refer the interested reader to [106] for a further overview of related results. We note that for several results which we will state, it is not required that the r.v.s be identically distributed, although we only state

the results for that setting.

Lemma 2.16 ([105] Theorem 2)**.** *Suppose that for some $p \geq 2$, $\{X_i, i \geq 1\}$ is a collection of i.i.d. zero-mean r.v.s. s.t. $\mathbb{E}[|X_1|^p] < \infty$. Then for all $k \geq 1$,*

$$\mathbb{E}\big[\big|\sum_{i=1}^{k} X_i\big|^p\big] \leq (5p)^p \mathbb{E}[|X_1|^p] k^{\frac{p}{2}}.$$

For later use, here we also state a result similar to Lemma 2.16, but under different assumptions, e.g. requiring the r.v.s be non-negative but not necessarily centered, and only requiring finite first moment. The result follows immediately from [107] Theorem 2.5.

Lemma 2.17 ([107] Theorem 2.5)**.** *Suppose that for some $p \geq 1$, $\{X_i, i \geq 1\}$ is a collection of i.i.d. non-negative r.v.s. s.t. $\mathbb{E}[X_1^p] < \infty$. Then for all $k \geq 1$,*

$$\mathbb{E}\big[\big(\sum_{i=1}^{k} X_i\big)^p\big] \leq (2p)^p \max\big((k\mathbb{E}[X_1])^p, k\mathbb{E}[X_1^p]\big).$$

With Corollary 2.8 and Lemma 2.16 in hand, we now complete the proof of Lemma 2.11.

Proof of Lemma 2.11. Applying Lemma 2.16 with $X_i = N_i(t) - t$, we find that

$$\mathbb{E}\big[\big|\sum_{i=1}^{k} N_i(t) - kt\big|^r\big] \leq (5r)^r E\big[|N_1(t) - t|^r\big] k^{\frac{r}{2}}.$$

Combining with Corollary 2.8 and some straightforward algebra completes the proof. □

2.5.2 Bounding the central moments of $\sum_{i=1}^{k} N_i(t)$ for $t \in [0, 1]$

In thus subsection we bound the central moments of $\sum_{i=1}^{k} N_i(t)$ for $t \in [0, 1]$. We will use an argument similar to that used in the proof of [63] Lemma 5. However, in contrast to the arguments of [63], here all quantities are made completely explicit.

Lemma 2.18. *Suppose that* $\mathbb{E}[S] = 1$. *Then for all* $k \geq 1, p \geq 2, t \in [0,1]$, *and* $\theta > 0$,

$$\mathbb{E}\Big[\big|\sum_{i=1}^{k} N_i(t) - kt\big|^p\Big] \leq \exp(\theta)\big(\frac{10^5 p^4}{1 - \mathbb{E}[\exp(-\theta S)]}\big)^{p+2} \max\big(kt, (kt)^{\frac{p}{2}}\big). \tag{2.34}$$

Our proof proceeds by first proving a somewhat weaker bound, and then leveraging this bound to prove the desired result.

A useful weaker bound

We now establish the aforementioned weaker bound, which we will ultimately use to prove Lemma 2.18. Intuitively, this weaker bound follows by "interpreting" $\sum_{i=1}^{k} N_i(t)$ as a type of "modified binomial" distribution, where each renewal process has "a success probability" of having had at least one event. We note that this weaker bound, and its proof, are similar to that of Lemma 9 in [63] (although in [63] the corresponding results are non-explicit). In particular, we prove the following.

Lemma 2.19. *Suppose that* $\mathbb{E}[S] = 1$. *Then for all* $k \geq 1, p \geq 2, t \in [0,1]$, *and* $\theta > 0$,

$$\mathbb{E}\big[\big|\sum_{i=1}^{k} N_i(t) - kt\big|^p\big] \leq \exp(\theta)\big(\frac{10^3 p^3}{1 - \mathbb{E}[\exp(-\theta S)]}\big)^{p+2} \max\big(kt, (kt)^p\big). \tag{2.35}$$

Proof of Lemma 2.19. We note that here it is important to correctly capture the joint scaling of k and t, so e.g. a naive application of Lemma 2.16 will not suffice. Instead, we proceed as follows. It follows from (2.20) that the left-hand-side of (2.35) is at most

$$\mathbb{E}\big[\big(\sum_{i=1}^{k} N_i(t) + kt\big)^p\big] \;\leq\; 2^{p-1}\big(\mathbb{E}\big[\big(\sum_{i=1}^{k} N_i(t)\big)^p\big] + (kt)^p\big). \tag{2.36}$$

We now bound the term $\mathbb{E}\big[\big(\sum_{i=1}^{k} N_i(t)\big)^p\big]$ appearing in (2.36). Let us fix some $t \in [0,1]$, and let $\{B_i, i \geq 1\}$ denote a sequence of i.i.d. Bernoulli r.v. s.t $\mathbb{P}(B_i = 1) = p_t \stackrel{\Delta}{=} \mathbb{P}(R(S) \leq t)$, and $\mathbb{P}(B_i = 0) = 1 - p_t$. Note that we may construct $\{N_i(t), i \geq 1\}$, $\{N_{o,i}(t), i \geq 1\}, \{B_i, i \geq 1\}$ on the same probability space s.t. w.p.1 $N_i(t) \leq B_i\big(1 +$

$N_{o,i}(t)\big)$ for all $i \geq 1$, with $\{N_{o,i}(t), i \geq 1\}, \{B_i, i \geq 1\}$ mutually independent. Let $M_t \stackrel{\Delta}{=} \sum_{i=1}^{k} B_i$. Then it follows from Lemma 2.17, Corollary 2.7, the fact that $t \leq 1$, and Jensen's inequality that

$$\begin{aligned}
\mathbb{E}\big[\big(\sum_{i=1}^{k} N_i(t)\big)^p\big] &\leq \mathbb{E}\big[\Big(\sum_{i=1}^{M_t}\big(1+N_{o,i}(t)\big)\Big)^p\big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[\Big(\sum_{i=1}^{M_t}\big(1+N_{o,i}(t)\big)\Big)^p|M_t\big]\Big] \\
&\leq \mathbb{E}\Big[(2p)^p \max\Big(\big(M_t\mathbb{E}[1+N_o(t)]\big)^p, M_t\mathbb{E}[\big(1+N_o(t)\big)^p]\Big)\Big] \\
&\leq (2p)^p\mathbb{E}[\big(1+N_o(t)\big)^p]\mathbb{E}\big[\max\big(M_t^p, M_t\big)\big] \\
&= (2p)^p\mathbb{E}[\big(1+N_o(t)\big)^p]\mathbb{E}[M_t^p] \\
&\leq (2p)^p\mathbb{E}[\big(1+N_o(1)\big)^p]\mathbb{E}[M_t^p] \\
&\leq \exp(\theta)\big(\frac{200p^2}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{p+2}\mathbb{E}[M_t^p].
\end{aligned}$$

Further applying Lemma 2.17 to conclude that

$$\mathbb{E}[M_t^p] = \mathbb{E}\big[\big(\sum_{i=1}^{k} B_i\big)^p\big] \leq (2p)^p \max\big((kp_t)^p, kp_t\big),$$

we may combine the above and find that

$$\mathbb{E}\big[\big(\sum_{i=1}^{k} N_i(t)\big)^p\big] \leq \exp(\theta)\big(\frac{400p^3}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{p+2}\max\big(kp_t, (kp_t)^p\big).$$

Since it follows from the definition of the equilibrium distribution and p_t that $p_t \leq t$, we conclude that the left-hand-side of (2.35) is at most

$$2^{p-1}\Big(\exp(\theta)\big(\frac{400p^3}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{p+2}\max\big(kp_t, (kp_t)^p\big) + (kt)^p\Big).$$

Further combining with some straightforward algebra completes the proof. □

Proof of Lemma 2.18

We now use Lemma 2.19 to complete the proof of the desired result Lemma 2.18, proceeding by a case analysis. We note that a similar, albeit non-explicit, analysis appeared in [63].

Proof of Lemma 2.18. Let us fix some $t \in [0,1]$. We proceed by a case analysis. First, suppose $t \leq \frac{1}{k}$. In this case $\max\big(kt, (kt)^p\big) = kt$, and the desired result follows from Lemma 2.19.

Next, suppose $t \in (\frac{1}{k}, \frac{2}{k}]$. In this case,

$$\max\big(kt, (kt)^p\big) \quad = \quad (kt)^p \quad \leq \quad 2^{\frac{p}{2}}(kt)^{\frac{p}{2}},$$

and the result then follows from Lemma 2.19.

Alternatively, suppose $t \in \big(\frac{2}{k}, 1\big]$. Let $n_1(t) \stackrel{\Delta}{=} \lfloor kt \rfloor$. Noting that $t \geq \frac{2}{k}$ implies $n_1(t) > 0$, in this case we may define $n_2(t) \stackrel{\Delta}{=} \lfloor \frac{k}{n_1(t)} \rfloor$. Then the left-hand-side of (2.34) equals

$$\begin{aligned} & \mathbb{E}\big[\big| \sum_{m=1}^{n_1(t)} \sum_{l=1}^{n_2(t)} \big(N_{(m-1)n_2(t)+l}(t) - t\big) + \sum_{l=n_1(t)n_2(t)+1}^{k} \big(N_l(t) - t\big)\big|^p\big] \\ \leq \quad & 2^{p-1}\mathbb{E}\big[\big| \sum_{m=1}^{n_1(t)} \sum_{l=1}^{n_2(t)} \big(N_{(m-1)n_2(t)+l}(t) - t\big)\big|^p\big] \qquad (2.37) \\ & +2^{p-1}\mathbb{E}\big[\big| \sum_{l=n_1(t)n_2(t)+1}^{k} \big(N_l(t) - t\big)\big|^p\big] \quad \text{by (2.20).} \qquad (2.38) \end{aligned}$$

We now bound (2.37). It follows from Lemma 2.16 that (2.37) is at most

$$\begin{aligned} & (10p)^p \big(n_1(t)\big)^{\frac{p}{2}} \mathbb{E}\big[\big|\sum_{l=1}^{n_2(t)} \big(N_l(t)-t\big)\big|^p\big] \\ \leq \;& (10p)^p \big(n_1(t)\big)^{\frac{p}{2}} \exp(\theta)\big(\frac{10^3 p^3}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{p+2} \max\Big(tn_2(t), \big(tn_2(t)\big)^p\Big), \end{aligned} \tag{2.39}$$

the final inequality following from Lemma 2.19. We now bound the term $tn_2(t)$ appearing in (2.39). In particular,

$$tn_2(t) \;=\; t\big\lfloor \frac{k}{\lfloor kt \rfloor} \big\rfloor \;\leq\; \frac{kt}{kt-1}. \tag{2.40}$$

But since $t \geq \frac{2}{k}$ implies $kt \geq 2$, and $g(z) \stackrel{\Delta}{=} \frac{z}{z-1}$ is a decreasing function of z on $(1,\infty)$, it follows from (2.40) that

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

Thus $\max\Big(tn_2(t), \big(tn_2(t)\big)^p\Big) \leq 2^p$. As in addition $n_1(t) \leq kt$, it then follows from (2.39) that (2.37) is at most

$$\exp(\theta)\big(\frac{2\times 10^4 p^4}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{p+2} (kt)^{\frac{p}{2}}. \tag{2.41}$$

We now bound (2.38). Note that the sum $\sum_{l=n_1(t)n_2(t)+1}^{k} \big(N_l(t)-\mu_S t\big)$ appearing in (2.38) is taken over $k - n_1(t)n_2(t)$ terms. Furthermore,

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

As $n_1(t) \leq kt$, it thus follows from Lemma 2.16 that (2.38) is at most

$$(10p)^p \times (kt)^{\frac{p}{2}} \times \mathbb{E}\big[|N_1(t)-t|^p\big]. \tag{2.42}$$

Further using Lemma 2.19 to bound $\mathbb{E}\big[|N_1(t) - t|^p\big]$ by $\exp(\theta)\big(\frac{10^3 p^3}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{p+2}$ (since $t \leq 1$), we conclude that (2.38) is at most

$$\exp(\theta)\big(\frac{10^4 p^4}{1 - \mathbb{E}[\exp(-\theta S)]}\big)^{p+2}(kt)^{\frac{p}{2}}. \tag{2.43}$$

Using (2.41) to bound (2.37) and (2.43) to bound (2.38) completes the proof. □

2.5.3 Proof of main result Theorem 2.1

In this subsection we complete the proof of our main result Theorem 2.1. We proceed by using Lemmas 2.11 and 2.18 to prove that the conditions of Theorem 2.4 are met, in conjunction with the stochastic comparison results Theorems 2.2 and 2.3. Before completing the proof, we provide one additional auxiliary result, bounding the central moments of $\mu_A \sum_{i=1}^k A_i$. The proof follows immediately from Lemma 2.16, the easily verified (using (2.20)) fact that for all $r \geq 2$,

$$\mathbb{E}\big[|\mu_A A - 1|^r\big] \leq 2^{r-1}(\mathbb{E}[A^r]\mu_A^r + 1) \leq 2^r \mathbb{E}[A^r]\mu_A^r,$$

and some straightforward algebra, and we omit the details.

Lemma 2.20. *Suppose that $0 < \mu_A < \infty$, and $\mathbb{E}[A^{r_3}] < \infty$ for some $r_3 \geq 2$. Then for all $k \geq 1$,*

$$\mathbb{E}\big[|\mu_A \sum_{i=1}^k A_i - k|^{r_3}\big] \leq (10 r_3)^{r_3} \mathbb{E}[A^{r_3}]\mu_A^{r_3} k^{\frac{r_3}{2}}.$$

We now complete the proof of our main result Theorem 2.1. We proceed by first using Lemmas 2.11, 2.18, and 2.20 to verify that the conditions of Theorem 2.4 are met with particular constants. For ease of exposition, we state this result as its own stand-alone lemma.

Lemma 2.21. *Suppose that $0 < \mu_A, \mu_S < \infty$, and that $\mathbb{E}[S^r] < \infty, \mathbb{E}[A^r] < \infty$ for some $r > 2$.*

Proof. It follows from Lemmas 2.11, 2.18, and 2.20 that for each integer $n' \geq 1$ s.t. $n' > \mu_A$, the conditions of Theorem 2.4 are met with the following parameters:

$$r_1 = r_2 = r_3 = r \quad , \quad s_1 = s_3 = \frac{r}{2},$$

$$C_1 = \mathbb{E}[S^r] \exp(\theta) \big(\frac{10^8 r^3}{1 - \mathbb{E}[\exp(-\theta S)]} \big)^{r+2},$$

$$C_2 = \exp(\theta) \big(\frac{10^5 r^4}{1 - \mathbb{E}[\exp(-\theta S)]} \big)^{r+2},$$

$$C_3 = (10r)^r \mathbb{E}[A^r] \mu_A^r.$$

Thus, applying Theorem 2.4 and some straightforward algebra (e.g. the fact that $1 + C_i \leq 2C_i$ for $i = 1, 2, 3$), we find that for all $z \geq 16$, $\mathbb{P}\bigg(\sup_{t \geq 0} \Big(A(t) - \sum_{i=1}^{n'} N_i(t) \Big) \geq z \bigg)$ is at most

$$\begin{aligned}
8 \times & \left(\frac{10^6 (3r)^5}{(\frac{r}{2} - 1)^2 (\frac{r}{2})^2 (r-2)} \right)^{3r+1} \times \left(\mathbb{E}[S^r] \exp(\theta) \big(\frac{10^8 r^3}{1 - \mathbb{E}[\exp(-\theta S)]} \big)^{r+2} \right) \\
& \times \left(\exp(\theta) \big(\frac{10^5 r^4}{1 - \mathbb{E}[\exp(-\theta S)]} \big)^{r+2} \right) \times \left((10r)^r \mathbb{E}[A^r] \mu_A^r \right) \\
& \times \Bigg(n'^{\frac{r}{2}} (n' - \mu_{A,n'})^{-\frac{r}{2}} z^{-\frac{r}{2}} \\
& \quad + \quad n'^{\frac{r}{2}} (n' - \mu_{A,n'})^{-\frac{r}{2}} z^{-\frac{r}{2}} \\
& \quad + \quad (n' - \mu_{A,n'})^{-\frac{r}{2}} (n')^r \mu_{A,n'}^{-\frac{r}{2}} z^{-\frac{r}{2}} \Bigg),
\end{aligned}$$

which after some further straightforward algebra (and the fact that $n' > \mu_{A,n'}$) is itself bounded by

$$\mathbb{E}[S^r] \mathbb{E}[A^r] \mu_A^r \exp(2\theta) \Big(\frac{10^{23} r^8}{(r-2)^3 \big(1 - \mathbb{E}[\exp(-\theta S)]\big)} \Big)^{4r} \times \Big(\frac{n'}{\mu_{A,n'}} \Big)^{\frac{r}{2}} \times n'^{\frac{r}{2}} \times \Big(n' - \mu_{A,n'} \Big)^{-\frac{r}{2}} \times z^{-\frac{r}{2}}. \tag{2.44}$$

Next, observe that the definition of $\mu_{A,n'}$ ensures that

$$\frac{n'}{\mu_{A,n'}} \quad \leq \quad 2 \quad , \quad \text{and} \quad n' - \mu_{A,n'} \quad \geq \quad \frac{1}{2}(n' - \mu_A), \tag{2.45}$$

the second inequality following from the fact that if $\mu_A < \frac{1}{2}n'$, then $n' - \mu_{A,n'} = \frac{1}{2}n' \geq \frac{1}{2}(n' - \mu_A)$; while if $\mu_A \geq \frac{1}{2}n'$, then $\mu_A = \mu_{A,n'}$, and hence $\frac{1}{2}(n' - \mu_A) = \frac{1}{2}(n' - \mu_{A,n'}) \leq n' - \mu_{A,n'}$. Combining (2.44) and (2.45) with some straightforward algebra, we conclude that for all $n' \geq 1$ s.t. $\rho_{n'} < 1$, all $z \geq 16$, and all $\theta > 0$, $\mathbb{P}\Big(\sup_{t \geq 0} \Big(A(t) - \sum_{i=1}^{n'} N_i(t) \Big) \geq z \Big)$ is at most

$$\mathbb{E}[S^r]\mathbb{E}[A^r]\mu_A^r \exp(2\theta) \Big(\frac{10^{24} r^8}{(r-2)^3 \big(1 - \mathbb{E}[\exp(-\theta S)]\big)} \Big)^{4r} \times \Big(z(1 - \rho_{n'}) \Big)^{-\frac{r}{2}}. \tag{2.46}$$

Further noting that $z \in (0, 16)$ implies (2.49) is at least one (by a straightforward calculation), we conclude that for all $n' \geq 1$ s.t. $\rho_{n'} < 1$, and all $z > 0$, $\mathbb{P}\Big(\sup_{t \geq 0} \Big(A(t) - \sum_{i=1}^{n'} N_i(t) \Big) \geq z \Big)$ is at most (2.49). We next show how to get rid of the term $1 - \mathbb{E}[\exp(-\theta S)]$ by an appropriate choice of θ. Note that for all $\theta > 0$, w.p.1, $\exp(\theta S) \geq 1 + \theta S$, and hence

$$\begin{aligned} \exp(-\theta S) &\leq \frac{1}{1 + \theta S} \\ &= 1 - \theta S + \frac{\theta^2 S^2}{1 + \theta S} \\ &\leq 1 - \theta S + \theta^2 S^2. \end{aligned}$$

It follows that for all $\theta > 0$,

$$1 - \mathbb{E}[\exp(-\theta S)] \geq \theta \mathbb{E}[S] - \theta^2 \mathbb{E}[S^2]. \tag{2.47}$$

Taking $\theta = \frac{\mathbb{E}[S]}{2\mathbb{E}[S^2]}$ and recalling that $\mathbb{E}[S] = 1$, we find that

$$\frac{1}{1 - \mathbb{E}[\exp(-\frac{1}{2\mathbb{E}[S^2]}S)]} \leq 4\mathbb{E}[S^2] \leq 4\big(\mathbb{E}[S^r]\big)^{\frac{2}{r}}, \tag{2.48}$$

the final inequality following from Jensen's inequality and the fact that $r \geq 2$. Further noting that $\mathbb{E}[S] = 1$ implies $\exp(2\frac{\mathbb{E}[S]}{2\mathbb{E}[S^2]}) \leq 3$, we may combine the above with (2.49) to conclude that for all $n' \geq 1$ s.t. $\rho_{n'} < 1$, and all $z \geq 16$, $\mathbb{P}\Bigg(\sup_{t\geq 0} \left(A(t) - \sum_{i=1}^{n'} N_i(t) \right) \geq z \Bigg)$ is at most

$$\big(\mathbb{E}[S^r]\big)^3 \mathbb{E}[A^r] \mu_A^r \Big(10^{26} r^8 (r-2)^{-3}\Big)^{4r} \times \Big(z(1-\rho_{n'})\Big)^{-\frac{r}{2}}. \tag{2.49}$$

Letting $n' = n$ and plugging in $z = \frac{x}{1-\rho_n}$ completes the proof of the first part of the theorem when $\mathbb{E}[S] = 1$, and the general result follows by rescaling A and S by $\mathbb{E}[S]$.

To prove the second part, let us plug in $n' = n - \lfloor \frac{1}{2}(n - \mu_A) \rfloor$ and $z = \lfloor \frac{1}{2}(n - \mu_A) \rfloor$. First, let us treat the case that $\rho_n \leq 1 - \frac{4}{n}$, which implies that

$$\frac{1}{4}(n - \mu_A) \leq \lfloor \frac{1}{2}(n - \mu_A) \rfloor \leq \frac{1}{2}(n - \mu_A),$$

from which we conclude (after some straightforward algebra) that $z(1 - \rho_{n'})$ is at least

$$\begin{aligned}
&\big(\frac{1}{4}(n - \mu_A)\big) \times \left(1 - \frac{\mu_A}{\left(n - \left(\frac{1}{2}(n - \mu_A)\right)\right)}\right) \\
&= \frac{1}{4}\frac{(n - \mu_A)^2}{n + \mu_A} \\
&\geq \frac{1}{8}\frac{(n - \mu_A)^2}{n} \\
&= \frac{n}{8}(1 - \rho_n)^2.
\end{aligned}$$

Combining with some straightforward algebra proves that, if $\mathbb{E}[S] = 1$, then for all $n \geq 5$

s.t. $\rho_n \leq 1 - \frac{4}{n}$, the s.s.p.d. is at most

$$\left(\mathbb{E}[S^r]\right)^3 \mathbb{E}[A^r] \mu_A^r \left(10^{27} r^8 (r-2)^{-3}\right)^{4r} \times \left(n(1-\rho_n)^2\right)^{-\frac{r}{2}}. \tag{2.50}$$

Noting that 1: if $n \geq 5$ and $\rho_n > 1 - \frac{4}{n}$ then (2.50) is at least one, and 2: if $n \leq 4$ then (2.50) is at least one, and again rescaling A and S by $\mathbb{E}[S]$ (to reduce to the setting in which $\mathbb{E}[S] = 1$) completes the proof of the second part of the theorem. □

2.6 Conclusion

In this chapter, we proved the first simple and explicit bounds for general $GI/GI/n$ queues which scale universally as $\frac{1}{1-\rho}$, across both the classical and Halfin-Whitt heavy-traffic regimes. Our bounds are very simple functions of only a single moment of the inter-arrival and service time distributions, where the strength of our bounds (e.g. in the form of tail decay rate) depends on the order of this given moment. We also provide the first bounds of this kind for the steady-state probability of delay, which provides new insights into the behavior of queues in the Halfin-Whitt regime.

Our results leave many interesting directions for future research. First, there is the task of tightening our bounds, to make them practically useful. In essentially all cases we opted for simplicity over tightening constants, so a careful pass through essentially the same analysis may already go a long ways here. It is also plausible that an alternative analysis of the relevant supremum could yield considerably tighter bounds. If this "bridge to practicality" were acheived, the corresponding results would essentially be as powerful as Kingman's bound, but in the multi-server setting, and potentially quite impactful both in theory and practice.

Second, it would be very interesting to connect our bounds to the asymptotic results of [63, 83]. Namely, the results of [63, 83] suggest that at least asymptotically (in the Halfin-Whitt regime), stronger bounds should be possible. Whether these stronger asymptotic

results can be made into simple, explicit, non-asymptotic bounds remains an interesting open question, where we note that related questions have been discussed in [47].

Third, it would also be interesting to connect our bounds to the results of [55, 56], which (in many settings) give necessary and sufficient conditions for the steady-state queue length to have finite rth moment. The results of [55, 56] actually show that having more servers leads to more moments being finite in a subtle way, and e.g. shows that in the Halfin-Whitt regime, the number of moments which are finite grows as the numbers of servers diverges. Such a result is in some ways considerably stronger than our own results in this regard. However, our results also speak explicitly to certain moments scaling gracefully with $\frac{1}{1-\rho}$, which the results of [55, 56] do not speak to. Understanding this disconnect, and the connection between finiteness of moments and the scaling of those moments is an interesting open question. This matter also relates to another open question: understanding whether related bounds are possible when service times are heavy-tailed (i.e. infinite variance), for which essentially nothing is known in the Halfin-Whitt regime. Indeed, there may be quite subtle interactions between which moments of the service time distribution exist, which moments of the steady-state queue-length exist, and how those moments scale, where we note that related questions have been previously studied in the single-server setting [108]. Also, one may ask whether a simplified analysis, yielding stronger bounds, is possible if one assumes that all moments of the service time distribution are finite (e.g. in the case of phase-type service times).

Fourth, it is natural to extend our results to other queueing settings (in addition to that of heavy-tails), e.g. queues with abandonments, networks of queues, etc. We note that the results of [63] also give bounds for the transient setting, and could be naturally extended to consider time-varying arrival processes, another interesting direction. One could also attempt to derive simple, explicit, and universal bounds for other quantities of interest in the analysis of queues, e.g. the rate of convergence to stationarity, where we refer the reader to [109] for some relevant discussion in the Markovian setting. It may also be interesting

to combine our bounds with the robust optimization approach of [91], perhaps yielding stronger bounds for multi-server queues within that framework.

Fifth, as mentioned previously, there are heavy-traffic settings in which $\frac{1}{1-\rho}$ is not the correct scaling, e.g. when $\rho \uparrow 1$ but the steady-state probability of delay $\to 0$. Refining the bounds to scale correctly in this setting as well, so as to be truly universal, also remains an interesting direction for future research.

On a final note, and taking a broader view of the literature on queueing theory, there is the meta-question of how to conceptualize the trade-off between simplicity/explicitness, and accuracy, in approximations for multi-server queues. This question is particularly interesting in the Halfin-Whitt regime, where the inherent complexity of the weak limits that arise brings this question front and center. The following are but a few interesting questions along these lines. What is the right notion of "complexity" in queueing approximations? How should one compare analytical bounds with results derived from simulation and numerical procedures? What is the formal algorithmic complexity of both numerical computation, and simulation, for the limiting processes which arise? And last, but by no means least, which types of approximations may be most useful in practice?

2.7 Appendix

2.7.1 Proof of Theorem 2.3

We begin by citing the relevant result of [83].

Lemma 2.22 ([83] Theorem 4)**.** *Under the same assumptions as Theorem 2.2, for all* $x \geq 0$*,* $\mathbb{P}\big(Q^n(\infty) \geq x\big)$ *is at most*

$$\inf_{\substack{\delta \geq 0 \\ \eta \in [0,n]}} \mathbb{P}\Bigg(\max\Bigg(\sup_{0 \leq t \leq \delta} \big(A(t) - \sum_{i=1}^{\eta} N_i(t)\big) \ , \ \sup_{t \geq \delta} \big(A(t) - \sum_{i=1}^{n} N_i(t)\big) + \sum_{i=\eta+1}^{n} N_i(\delta) \Bigg) + \sum_{i=\eta+1}^{n} I(N_i(\delta) = 0) \ \geq \ x - \eta \Bigg).$$

With Lemma 2.22 in hand, we now complete the proof of Theorem 2.3.

Proof of Theorem 2.3. First, let us prove that $n\mu_S > \mu_A$ implies

$$\left(n - \lfloor \frac{1}{2}(n - \frac{\mu_A}{\mu_S})\rfloor\right)\mu_S > \mu_A. \tag{2.51}$$

Indeed,

$$\begin{aligned}\left(n - \lfloor \frac{1}{2}(n - \frac{\mu_A}{\mu_S})\rfloor\right)\mu_S &\geq \left(n - \frac{1}{2}(n - \frac{\mu_A}{\mu_S})\right)\mu_S \\ &= \frac{1}{2}(n\mu_S + \mu_A) \quad > \quad \mu_A.\end{aligned}$$

Next, taking $x = n, \eta = n - \lfloor \frac{1}{2}(n - \frac{\mu_A}{\mu_S})\rfloor$, we conclude from Lemma 2.22 that $\mathbb{P}\big(Q^n(\infty) \geq n\big)$ is at most

$$\begin{aligned}\inf_{\delta\geq 0}\mathbb{P}\Bigg(\max\Bigg(\sup_{0\leq t\leq\delta}\big(A(t) - \sum_{i=1}^{n-\lfloor \frac{1}{2}(n-\frac{\mu_A}{\mu_S})\rfloor} N_i(t)\big), \sup_{t\geq\delta}\big(A(t) - \sum_{i=1}^{n} N_i(t)\big) + \sum_{i=n-\lfloor \frac{1}{2}(n-\frac{\mu_A}{\mu_S})\rfloor+1}^{n} N_i(\delta)\Bigg) \\ + \sum_{i=n-\lfloor \frac{1}{2}(n-\frac{\mu_A}{\mu_S})\rfloor+1}^{n} I(N_i(\delta) = 0) \quad \geq \quad \lfloor \frac{1}{2}(n - \frac{\mu_A}{\mu_S})\rfloor\Bigg).\end{aligned}$$

Applying monotonicity and a union bound, we further find that for all $\epsilon, T > 0$, $\mathbb{P}\big(Q^n(\infty) \geq n\big)$ is at most

$$\mathbb{P}\left(\sup_{t\geq 0}\big(A(t) - \sum_{i=1}^{n-\lfloor \frac{1}{2}(n-\frac{\mu_A}{\mu_S})\rfloor} N_i(t)\big) \geq \lfloor \frac{1}{2}(n - \frac{\mu_A}{\mu_S})\rfloor - \epsilon \right) \tag{2.52}$$

$$+ \quad \mathbb{P}\left(\sup_{t\geq T}\big(A(t) - \sum_{i=1}^{n} N_i(t)\big) + \sum_{i=n-\lfloor \frac{1}{2}(n-\frac{\mu_A}{\mu_S})\rfloor+1}^{n} N_i(T) \geq 0 \right) \tag{2.53}$$

$$+ \quad \mathbb{P}\left(\sum_{i=1}^{n} I(N_i(T) = 0) \geq \epsilon \right). \tag{2.54}$$

We next bound (2.53), which by stationary increments and another union bound is at most

$$\mathbb{P}\left(A(T) - \sum_{i=1}^{n-\lfloor \frac{1}{2}(n-\frac{\mu_A}{\mu_S})\rfloor} N_i(T) \geq -T^{\frac{1}{2}} \right) + \mathbb{P}\left(\sup_{t\geq 0} \big(A(t) - \sum_{i=1}^{n} N_i(t)\big) \geq T^{\frac{1}{2}} \right).$$

It follows from the well-known Strong law of large numbers for renewal processes, and (2.51), that

$$\lim_{T\to\infty} \mathbb{P}\left(A(T) - \sum_{i=1}^{n-\lfloor \frac{1}{2}(n-\frac{\mu_A}{\mu_S})\rfloor} N_i(T) \geq -T^{\frac{1}{2}} \right) = 0 \ , \ \lim_{T\to\infty} \mathbb{P}\left(\sup_{t\geq 0} \big(A(t) - \sum_{i=1}^{n} N_i(t)\big) \geq T^{\frac{1}{2}} \right) = 0,$$

from which we conclude that

$$\lim_{T\to\infty} \mathbb{P}\left(\sup_{t\geq T} \big(A(t) - \sum_{i=1}^{n} N_i(t)\big) + \sum_{i=n-\lfloor \frac{1}{2}(n-\frac{\mu_A}{\mu_S})\rfloor+1}^{n} N_i(T) \geq 0 \right) = 0. \tag{2.55}$$

Furthermore, for all $\epsilon > 0$, it trivially holds that

$$\lim_{T\to\infty} \mathbb{P}\left(\sum_{i=1}^{n} I(N_i(T) = 0) \geq \epsilon \right) = 0. \tag{2.56}$$

Combining the above, we conclude that for all $\epsilon > 0$ (by taking the limit $T \to \infty$ above), $\mathbb{P}\big(Q^n(\infty) \geq n\big)$ is at most

$$\mathbb{P}\left(\sup_{t\geq 0} \big(A(t) - \sum_{i=1}^{n-\lfloor \frac{1}{2}(n-\frac{\mu_A}{\mu_S})\rfloor} N_i(t)\big) \geq \lfloor \frac{1}{2}(n - \frac{\mu_A}{\mu_S})\rfloor - \epsilon \right). \tag{2.57}$$

Letting $\epsilon \downarrow 0$ and applying continuity completes the proof. □

2.7.2 Proof of Lemma 2.2

We begin by citing the relevant result of [98].

Lemma 2.23 ([98] Theorem 2)**.** *Let $\{X_l, 1 \leq l \leq L\}$ be a completely general sequence of*

r.v.s. Suppose there exist $\nu > 0$, $\gamma > 1$, *and* $C > 0$ *such that for all* $\lambda > 0$ *and non-negative integers* $1 \leq i \leq j \leq L$, *it holds that*

$$\mathbb{P}\big(|\sum_{k=i}^{j} X_k| \geq \lambda\big) \leq \big(C(j-i+1)\big)^{\gamma}\lambda^{-\nu}.$$

Then it must also hold that

$$\mathbb{P}\big(\max_{i\in[1,L]}|\sum_{k=1}^{i} X_k| \geq \lambda\big) \leq 2^{\gamma}\bigg(1+\big(2^{-\frac{1}{\nu+1}}-2^{-\frac{\gamma}{\nu+1}}\big)^{-(\nu+1)}\bigg)(CL)^{\gamma}\lambda^{-\nu}.$$

With Lemma 2.23 in hand, we now complete the proof of Lemma 2.2.

Proof of Lemma 2.2. As the conditions of Lemma 2.23 and Lemma 2.2 are identical, it suffices to prove that $\nu \geq \gamma$ (along with the other assumptions of Lemma 2.23) implies

$$2^{\gamma}\bigg(1+\big(2^{-\frac{1}{\nu+1}}-2^{-\frac{\gamma}{\nu+1}}\big)^{-(\nu+1)}\bigg) \leq \big(6\frac{\nu+1}{\gamma-1}\big)^{\nu+1}. \tag{2.58}$$

Note that

$$2^{-\frac{1}{\nu+1}}-2^{-\frac{\gamma}{\nu+1}} = 2^{-\frac{1}{\nu+1}}\big(1-2^{-\frac{\gamma-1}{\nu+1}}\big). \tag{2.59}$$

As our assumptions imply $0 < \frac{\gamma-1}{\nu+1} < 1$, and it is easily verified that $1-2^{-z} \geq \frac{z}{2}$ for all $z \in [0,1]$, we conclude that

$$\big(1-2^{-\frac{\gamma-1}{\nu+1}}\big)^{-(\nu+1)} \leq \big(2\frac{\nu+1}{\gamma-1}\big)^{\nu+1}. \tag{2.60}$$

Combining (2.59) and (2.60) with the fact that (by our assumptions) $(\frac{\nu+1}{\gamma-1})^{\nu+1} \geq 1$ and

$2^\gamma \leq 2^\nu$, it follows that the left-hand-side of (2.58) is at most

$$\begin{aligned} & 2^\gamma \bigg(1 + 2(2\frac{\nu+1}{\gamma-1})^{\nu+1} \bigg). \\ \leq \quad & 3 \times 2^\gamma \times (2\frac{\nu+1}{\gamma-1})^{\nu+1} \\ \leq \quad & 6 \times 4^\nu \times (\frac{\nu+1}{\gamma-1})^{\nu+1} \\ \leq \quad & (6\frac{\nu+1}{\gamma-1})^{\nu+1}, \end{aligned}$$

completing the proof. □

2.7.3 Proof of Lemma 2.5

Proof of Lemma 2.5. Note that for $\lambda > 0$, $\mathbb{P}\bigg(\sup_{t \geq 0} \big(\phi(t) - \nu t \big) \geq \lambda \bigg)$ equals

$$\begin{aligned} & \mathbb{P}\bigg(\bigg(\bigcup_{k=0}^{\infty} \big\{ \phi(t) - \nu t \geq \lambda \text{ for some } t \in [2^k, 2^k + 1] \big\} \bigg) \bigcup \bigg\{ \phi(t) - \nu t \geq \lambda \text{ for some } t \in [0,1] \bigg\} \bigg) \\ \leq \quad & \sum_{k=0}^{\infty} \mathbb{P}\bigg(\sup_{t \in [2^k, 2^{k+1}]} \big(\phi(t) - \nu t \big) \geq \lambda \bigg) \qquad (2.61) \\ & + \quad \mathbb{P}\bigg(\sup_{t \in [0,1]} \big(\phi(t) - \nu t \big) \geq \lambda \bigg). \qquad (2.62) \end{aligned}$$

We now bound (2.61), and proceed by bounding (for each $k \geq 0$)

$$\mathbb{P}\bigg(\sup_{t \in [2^k, 2^{k+1}]} \big(\phi(t) - \nu t \big) \geq \lambda \bigg). \qquad (2.63)$$

Since $t \in [2^k, 2^{k+1}]$ implies $\nu t \geq \nu 2^k$, we conclude that (2.63) is at most

$$\mathbb{P}\bigg(\sup_{t \in [2^k, 2^{k+1}]} \phi(t) \geq \lambda + \nu 2^k \bigg),$$

which by adding and subtracting $\phi(2^k)$, and applying stationary increments and a union bound, is at most

$$\begin{aligned}
&\mathbb{P}\Big(\big(\sup_{t\in[2^k,2^{k+1}]} \phi(t) - \phi(2^k)\big) + \phi(2^k) \geq \lambda + \nu 2^k \Big) \\
&\leq \quad \mathbb{P}\Big(\sup_{t\in[2^k,2^{k+1}]} \phi(t) - \phi(2^k) \geq \frac{1}{2}(\lambda + \nu 2^k) \Big) + \mathbb{P}\Big(\phi(2^k) \geq \frac{1}{2}(\lambda + \nu 2^k) \Big) \\
&= \quad \mathbb{P}\Big(\sup_{t\in[0,2^k]} \phi(t) \geq \frac{1}{2}(\lambda + \nu 2^k) \Big) + \mathbb{P}\Big(\phi(2^k) \geq \frac{1}{2}(\lambda + \nu 2^k) \Big) \\
&\leq \quad 2\mathbb{P}\Big(\sup_{t\in[0,2^k]} \phi(t) \geq \frac{1}{2}(\lambda + \nu 2^k) \Big). \qquad (2.64)
\end{aligned}$$

We proceed to bound (2.64) by breaking the supremum into two parts, one part taken over integer points, one part taken over intervals of length one corresponding to the regions between these integer points. In particular, the assumptions of the lemma, combined with a union bound and stationary increments, ensure that

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{t\in[0,2^k]} \phi(t) \geq \frac{1}{2}(\lambda + \nu 2^k) \Big) \\
&\leq \quad \mathbb{P}\Big(\sup_{j\in\{0,\dots,2^k\}} \phi(j) + \sup_{\substack{j\in\{0,\dots,2^k-1\} \\ t\in[0,1]}} \big(\phi(j+t) - \phi(j)\big) \geq \frac{1}{2}(\lambda + \nu 2^k) \Big) \\
&\leq \quad \mathbb{P}\Big(\sup_{j\in\{0,\dots,2^k\}} \phi(j) \geq \frac{1}{4}(\lambda + \nu 2^k) \Big) + 2^k \mathbb{P}\Big(\sup_{t\in[0,1]} \phi(t) \geq \frac{1}{4}(\lambda + \nu 2^k) \Big) \\
&\leq \quad \frac{H_1 4^{r_1} 2^{ks}}{(\lambda + \nu 2^k)^{r_1}} + \frac{H_2 4^{r_2} 2^k}{(\lambda + \nu 2^k)^{r_2}}, \qquad (2.65)
\end{aligned}$$

where the final inequality is applicable since $\lambda \geq 4Z$ implies $\frac{1}{4}(\lambda + \nu 2^k) \geq Z$, in which case the inequality follows from our assumptions. Combining (2.64) and (2.65), we conclude that (2.61) is at most

$$2\sum_{k=0}^{\infty} \frac{H_1 4^{r_1} 2^{ks}}{(\lambda + \nu 2^k)^{r_1}} + 2\sum_{k=0}^{\infty} \frac{H_2 4^{r_2} 2^k}{(\lambda + \nu 2^k)^{r_2}}. \qquad (2.66)$$

We now treat two cases. First, suppose $\lambda > \nu$. Then (2.66) is at most

$$\begin{aligned}
& 2H_1 4^{r_1} \sum_{k=0}^{\lceil \log_2(\frac{\lambda}{\nu}) \rceil - 1} \frac{2^{ks}}{\lambda^{r_1}} \\
& \quad + \quad 2H_2 4^{r_2} \sum_{k=0}^{\lceil \log_2(\frac{\lambda}{\nu}) \rceil - 1} \frac{2^{k}}{\lambda^{r_2}} \\
& \quad + \quad 2H_1 4^{r_1} \sum_{k=\lceil \log_2(\frac{\lambda}{\nu}) \rceil}^{\infty} \frac{2^{-(r_1 - s)k}}{\nu^{r_1}} \\
& \quad + \quad 2H_2 4^{r_2} \sum_{k=\lceil \log_2(\frac{\lambda}{\nu}) \rceil}^{\infty} \frac{2^{-(r_2 - 1)k}}{\nu^{r_2}} \\
= \quad & 2H_1 4^{r_1} \lambda^{-r_1} \frac{2^{\lceil \log_2(\frac{\lambda}{\nu}) \rceil s} - 1}{2^s - 1} \\
& \quad + \quad 2H_2 4^{r_2} \lambda^{-r_2} (2^{\lceil \log_2(\frac{\lambda}{\nu}) \rceil} - 1) \\
& \quad + \quad 2H_1 4^{r_1} \nu^{-r_1} \frac{2^{-(r_1 - s)\lceil \log_2(\frac{\lambda}{\nu}) \rceil}}{1 - 2^{-(r_1 - s)}} \\
& \quad + \quad 2H_2 4^{r_2} \nu^{-r_2} \frac{2^{-(r_2 - 1)\lceil \log_2(\frac{\lambda}{\nu}) \rceil}}{1 - 2^{-(r_2 - 1)}} \\
\leq \quad & 4H_1 4^{r_1} \lambda^{-r_1} (\frac{\lambda}{\nu})^s \\
& \quad + \quad 4H_2 4^{r_2} \lambda^{-r_2} \frac{\lambda}{\nu} \\
& \quad + \quad 2H_1 \big(1 - 2^{-(r_1 - s)}\big)^{-1} 4^{r_1} \nu^{-r_1} (\frac{\lambda}{\nu})^{-(r_1 - s)} \\
& \quad + \quad 2H_2 \big(1 - 2^{-(r_2 - 1)}\big)^{-1} 4^{r_2} \nu^{-r_2} (\frac{\lambda}{\nu})^{-(r_2 - 1)},
\end{aligned}$$

with the first line of the final inequality following from the fact that $2^{\lceil \log_2(\frac{\lambda}{\nu}) \rceil s} - 1 \leq 2^s (\frac{\lambda}{\nu})^s$ and $2^s - 1 \geq 2^{s-1}$. Combining with the fact that $r_2 > 2$ implies $\big(1 - 2^{-(r_2 - 1)}\big)^{-1} \leq 2$, we conclude that if $\lambda > \nu$, then (2.66) is at most

$$\begin{aligned}
& 6H_1 \big(1 - 2^{-(r_1 - s)}\big)^{-1} 4^{r_1} \lambda^{-(r_1 - s)} \nu^{-s} \\
& \quad + \quad 8H_2 4^{r_2} \lambda^{-(r_2 - 1)} \nu^{-1}.
\end{aligned} \tag{2.67}$$

Combining with the fact that $\lambda > \nu > 0$ and $r_2 > 2$ implies $\lambda^{-(r_2-1)}\nu^{-1} \leq (\lambda\nu)^{-\frac{r_2}{2}}$, we conclude that if $\lambda > \nu$, then (2.66) is at most

$$\begin{aligned} & 6H_1\big(1-2^{-(r_1-s)}\big)^{-1}4^{r_1}\lambda^{-(r_1-s)}\nu^{-s} \\ + \;& 8H_2 4^{r_2}(\lambda\nu)^{-\frac{r_2}{2}}. \end{aligned} \tag{2.68}$$

Alternatively, suppose $\lambda \leq \nu$. Then (2.66) is at most

(2.69)

$$\begin{aligned} & 2H_1 4^{r_1}\sum_{k=0}^{\infty}\frac{2^{-(r_1-s)k}}{\nu^{r_1}} \\ + \;& 2H_2 4^{r_2}\sum_{k=0}^{\infty}\frac{2^{-(r_2-1)k}}{\nu^{r_2}} \\ \leq \;& 2H_1 4^{r_1}\nu^{-r_1}\big(1-2^{-(r_1-s)}\big)^{-1} \\ + \;& 4H_2 4^{r_2}\nu^{-r_2} \\ \leq \;& 2H_1\big(1-2^{-(r_1-s)}\big)^{-1}4^{r_1}\lambda^{-(r_1-s)}\nu^{-s} \qquad (2.70) \\ + \;& 4H_2 4^{r_2}(\lambda\nu)^{-\frac{r_2}{2}}, \end{aligned}$$

the final inequality following from the fact that $\nu \geq \lambda, r_1 > s, r_2 > 2$ implies $\nu^{-r_1} \leq \lambda^{-(r_1-s)}\nu^{-s}$, and $\nu^{-r_2} \leq (\lambda\nu)^{-\frac{r_2}{2}}$. Next, we claim that $\big(1-2^{-(r_1-s)}\big)^{-1} \leq 2(1+\frac{1}{r_1-s})$. Indeed, first, suppose $r_1 - s < 1$. In this case, as it is easily verified that $1-2^{-z} \geq \frac{z}{2}$ for all $z \in (0,1)$, the result follows. Alternatively, if $r_1 - s \geq 1$, then $\big(1-2^{-(r_1-s)}\big)^{-1} \leq 2$, completing the proof. Combining with (2.68) and (2.70), and our assumptions, it follows that in all cases (2.66), and hence (2.61), is at most

$$(1+\frac{1}{r_1-s})4^{r_1+r_2+1}\bigg(H_1\nu^{-s}\lambda^{-(r_1-s)} + H_2(\lambda\nu)^{-\frac{r_2}{2}}\bigg). \tag{2.71}$$

We next bound (2.62). First, suppose $\lambda \geq \nu$. Then our assumptions (applied with $t_0 = 1$) imply that (2.62) is at most

$$H_2 \lambda^{-r_2} \leq H_2 (\lambda \nu)^{-\frac{r_2}{2}}. \tag{2.72}$$

Alternatively, suppose that $\lambda < \nu$. Then applying our assumptions with $t_0 = \frac{\lambda}{\nu}$, along with a union bound, we conclude that

$$\mathbb{P}\Big(\sup_{t \in [0,1]} \big(\phi(t) - \nu t \big) \geq \lambda \Big)$$

is at most

$$\mathbb{P}\Big(\sup_{t \in [0, \frac{\lambda}{\nu}]} \big(\phi(t) - \nu t \big) \geq \lambda \Big) \tag{2.73}$$

$$+ \quad \mathbb{P}\Big(\sup_{t \in [\frac{\lambda}{\nu}, 1]} \big(\phi(t) - \nu t \big) \geq \lambda \Big). \tag{2.74}$$

It follows from our assumptions that (2.73) is at most

$$\mathbb{P}\Big(\sup_{t \in [0, \frac{\lambda}{\nu}]} \phi(t) \geq \lambda \Big) \quad \leq \quad H_2 (\frac{\lambda}{\nu})^{\frac{r_2}{2}} \lambda^{-r_2} \quad = \quad H_2 (\lambda \nu)^{-\frac{r_2}{2}}. \tag{2.75}$$

We next bound (2.74), which by stationary increments, a union bound, and our assumptions is at most

$$
\begin{aligned}
&\mathbb{P}\Big(\sup_{t\in[\frac{\lambda}{\nu},1]} \big(\phi(\frac{\lambda}{\nu}) + \phi(t) - \phi(\frac{\lambda}{\nu}) - \nu(t-\frac{\lambda}{\nu})\big) \geq 2\lambda \Big) \\
&\leq \quad \mathbb{P}\big(\phi(\frac{\lambda}{\nu}) \geq \frac{1}{2}\lambda\big) \\
&\quad + \quad \mathbb{P}\Big(\sup_{t\in[\frac{\lambda}{\nu},1]} \big(\phi(t) - \phi(\frac{\lambda}{\nu}) - \nu(t-\frac{\lambda}{\nu})\big) \geq \frac{3}{2}\lambda \Big) \\
&= \quad \mathbb{P}\big(\phi(\frac{\lambda}{\nu}) \geq \frac{1}{2}\lambda\big) \\
&\quad + \quad \mathbb{P}\Big(\sup_{s\in[0,1-\frac{\lambda}{\nu}]} \big(\phi(s+\frac{\lambda}{\nu}) - \phi(\frac{\lambda}{\nu}) - \nu s\big) \geq \frac{3}{2}\lambda \Big) \\
&\leq \quad \mathbb{P}\Big(\sup_{t\in[0,\frac{\lambda}{\nu}]} \phi(t) \geq \frac{1}{2}\lambda \Big) \\
&\quad + \quad \mathbb{P}\Big(\sup_{s\in[0,1-\frac{\lambda}{\nu}]} \big(\phi(s+\frac{\lambda}{\nu}) - \phi(\frac{\lambda}{\nu}) - \nu s\big) \geq \frac{3}{2}\lambda \Big) \\
&\leq \quad H_2(\frac{\lambda}{\nu})^{\frac{r_2}{2}}(\frac{\lambda}{2})^{-r_2} \\
&\quad + \quad \mathbb{P}\Big(\sup_{s\in[0,1-\frac{\lambda}{\nu}]} \big(\phi(s) - \nu s\big) \geq \frac{3}{2}\lambda \Big) \\
&\leq \quad 2^{r_2} H_2(\lambda\nu)^{-\frac{r_2}{2}} \qquad (2.76) \\
&\quad + \quad \mathbb{P}\Big(\sup_{t\in[0,1]} \big(\phi(t) - \nu t\big) \geq \frac{3}{2}\lambda \Big). \qquad (2.77)
\end{aligned}
$$

Let us define $f(z) \stackrel{\Delta}{=} \mathbb{P}\Big(\sup_{t\in[0,1]} \big(\phi(t) - \nu t\big) \geq z \Big)$. Then using (2.75) to bound (2.73), and (2.76) - (2.77) to bound (2.74), we conclude that for all $z \in [2Z, \nu)$,

$$f(z) \leq 2^{r_2+1} H_2(z\nu)^{-\frac{r_2}{2}} + f(\frac{3}{2}z). \qquad (2.78)$$

Let $j^* \stackrel{\Delta}{=} \sup\{j \in Z^+ \ : \ (\frac{3}{2})^j\lambda < \nu\}$. Then it follows from (2.78) that for all $j \in [0, j^*]$,

$$f\big((\frac{3}{2})^j\lambda\big) \leq 2^{r_2+1} H_2(\lambda\nu)^{-\frac{r_2}{2}} \big((\frac{3}{2})^{\frac{r_2}{2}}\big)^{-j} + f\big((\frac{3}{2})^{j+1}\lambda\big). \qquad (2.79)$$

Combining (2.79) with a straightforward induction and our assumptions, and noting that f is a non-increasing function, we conclude that for all $\lambda \in [2Z, \nu)$,

$$\begin{aligned} f(\lambda) &\leq 2^{r_2+1} H_2 (\lambda\nu)^{-\frac{r_2}{2}} \sum_{j=0}^{j^*} \big((\frac{3}{2})^{\frac{r_2}{2}} \big)^{-j} + f(\nu) \\ &\leq 2^{r_2+1} H_2 (\lambda\nu)^{-\frac{r_2}{2}} \sum_{j=0}^{\infty} (\frac{3}{2})^{-j} + H_2 \nu^{-r_2} \\ &\leq 2^{r_2+3} H_2 (\lambda\nu)^{-\frac{r_2}{2}} + H_2 \nu^{-r_2} \\ &\leq 2^{r_2+4} H_2 (\lambda\nu)^{-\frac{r_2}{2}}, \end{aligned} \tag{2.80}$$

the final inequality following from the fact that by assumption $\nu > \lambda$ and thus $\nu^{-r_2} \leq (\lambda\nu)^{-\frac{r_2}{2}}$. Thus using (2.72) to bound (2.62) in the case $\lambda \geq \nu$, and (2.80) to bound (2.62) in the case $\lambda < \nu$, we conclude that in all cases (2.62) is at most

$$2^{r_2+4} H_2 (\lambda\nu)^{-\frac{r_2}{2}}. \tag{2.81}$$

Using (2.71) to bound (2.61), and (2.81) to bound (2.62), demonstrates that for all $\lambda \geq 4Z$, $\mathbb{P}\Big(\sup_{t \geq 0} \big(\phi(t) - \nu t \big) \geq \lambda \Big)$ is at most

$$(1 + \frac{1}{r_1 - s}) 4^{r_1+r_2+2} \bigg(H_1 \nu^{-s} \lambda^{-(r_1-s)} + H_2 (\lambda\nu)^{-\frac{r_2}{2}} \bigg),$$

completing the proof. □

2.7.4 Proof of Lemma 2.9

We note that the proof of Lemma 2.9 follows quite similarly to the proof of Lemma 2.5, and as such whenever possible we will refer back to the proof of Lemma 2.5 for specific technical steps, etc.

Proof of Lemma 2.9. For $\lambda > 0$, $\mathbb{P}\Big(\sup_{k\geq 0} \big(\phi(k) - \nu k\big) \geq \lambda \Big)$ equals

$$\begin{aligned} &\mathbb{P}\Big(\bigcup_{k=0}^{\infty} \big\{ \phi(i) - \nu i \geq \lambda \text{ for some } i \in [2^k, 2^{k+1}] \big\} \Big) \\ &\leq \quad \sum_{k=0}^{\infty} \mathbb{P}\Big(\sup_{i\in[2^k,2^{k+1}]} \big(\phi(i) - \nu i\big) \geq \lambda \Big). \end{aligned} \tag{2.82}$$

We now bound (2.82), and proceed by bounding (for each $k \geq 0$)

$$\mathbb{P}\Big(\sup_{i\in[2^k,2^{k+1}]} \big(\phi(i) - \nu i\big) \geq \lambda \Big). \tag{2.83}$$

Since $i \in [2^k, 2^{k+1}]$ implies $\nu i \geq \nu 2^k$, we conclude that (2.83) is at most

$$\mathbb{P}\Big(\sup_{i\in[2^k,2^{k+1}]} \phi(t) \geq \lambda + \nu 2^k \Big),$$

which by adding and subtracting $\phi(2^k)$, and applying stationary increments and a union bound exactly as in the proof of Lemma 2.5, as well as the assumptions of the lemma, is at most

$$\begin{aligned} &\mathbb{P}\Big(\big(\sup_{i\in[2^k,2^{k+1}]} \phi(i) - \phi(2^k)\big) + \phi(2^k) \geq \lambda + \nu 2^k \Big) \\ &\leq \quad \mathbb{P}\Big(\sup_{i\in[2^k,2^{k+1}]} \phi(i) - \phi(2^k) \geq \frac{1}{2}(\lambda + \nu 2^k) \Big) + \mathbb{P}\Big(\phi(2^k) \geq \frac{1}{2}(\lambda + \nu 2^k) \Big) \\ &= \quad \mathbb{P}\Big(\sup_{i\in[1,2^k]} \phi(i) \geq \frac{1}{2}(\lambda + \nu 2^k) \Big) + \mathbb{P}\Big(\phi(2^k) \geq \frac{1}{2}(\lambda + \nu 2^k) \Big) \\ &\leq \quad 2\mathbb{P}\Big(\sup_{i\in[1,2^k]} \phi(i) \geq \frac{1}{2}(\lambda + \nu 2^k) \Big) \\ &\leq \quad \frac{2H_3 4^{r_3} 2^{k s_3}}{(\lambda + \nu 2^k)^{r_3}}. \end{aligned}$$

We conclude that (2.82) is at most

$$2H_3 4^{r_3} \sum_{k=0}^{\infty} \frac{2^{ks_3}}{(\lambda + \nu 2^k)^{r_3}}. \tag{2.84}$$

As in the proof of Lemma 2.5, we now treat two cases, with each case largely mirroring the proof of Lemma 2.5. First, suppose $\lambda > \nu$. Then (2.84) is at most

$$\begin{aligned}
& 2H_3 4^{r_3} \sum_{k=0}^{\lceil \log_2(\frac{\lambda}{\nu}) \rceil - 1} \frac{2^{ks_3}}{\lambda^{r_3}} + 2H_3 4^{r_3} \sum_{k=\lceil \log_2(\frac{\lambda}{\nu}) \rceil}^{\infty} \frac{2^{-(r_3-s_3)k}}{\nu^{r_3}} \\
= \; & 2H_3 4^{r_3} \left(\frac{2^{\lceil \log_2(\frac{\lambda}{\nu}) \rceil s_3} - 1}{2^{s_3} - 1} \lambda^{-3_1} + \frac{2^{r_3}}{2^{r_3} - 2^{s_3}} 2^{-(r_3-s_3)\lceil \log_2(\frac{\lambda}{\nu}) \rceil} \nu^{-r_3} \right) \\
\leq \; & 8H_3 4^{r_3} \left(1 - 2^{-(r_3-s_3)}\right)^{-1} \nu^{-s_3} \lambda^{-(r_3-s_3)} \\
\leq \; & 16H_3 4^{r_3} (1 + \frac{1}{r_3 - s_3}) \nu^{-s_3} \lambda^{-(r_3-s_3)}.
\end{aligned} \tag{2.85}$$

Alternatively, suppose $\lambda \leq \nu$. Then (2.84) is at most

$$\begin{aligned}
& 2H_3 4^{r_3} \sum_{k=0}^{\infty} \frac{2^{-(r_3-s_3)k}}{\nu^{r_3}} \\
= \; & 2H_3 4^{r_3} \left(1 - 2^{-(r_3-s_3)}\right)^{-1} \nu^{-r_3} \\
\leq \; & 4H_3 4^{r_3} \left(1 + \frac{1}{r_3 - s_3}\right) \nu^{-r_3} \\
\leq \; & 4H_3 4^{r_3} \left(1 + \frac{1}{r_3 - s_3}\right) \nu^{-s_3} \lambda^{-(r_3-s_3)}.
\end{aligned} \tag{2.86}$$

Using (2.85) - (2.86) to bound (2.82) completes the proof. □

2.7.5 Proof of Lemma 2.15

Proof of Lemma 2.15. Note that for all $j \geq 1$ and $\theta > 0$,

$$\begin{aligned}
\mathbb{P}\big(N_o(1) \geq j\big) &= \mathbb{P}(\sum_{i=1}^{j} S_i \leq 1) \\
&= \mathbb{P}\bigg(\exp\big(-\theta \sum_{i=1}^{j} S_i\big) \geq \exp(-\theta)\bigg) \\
&\leq \exp(\theta) \times \mathbb{E}^j[\exp(-\theta S)] \quad \text{by Markov's inequality.}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[N_o^p(1)] &= \sum_{j=0}^{\infty} j^p \mathbb{P}\big(N_o(1) = j\big) \\
&\leq \sum_{j=0}^{\infty} j^p \mathbb{P}\big(N_o(1) \geq j\big) \\
&\leq \exp(\theta) \times \sum_{j=0}^{\infty} j^p \big(\mathbb{E}[\exp(-\theta S)]\big)^j \\
&\leq \exp(\theta) \times \bigg(1 + \int_1^{\infty} (x+1)^p \big(\mathbb{E}[\exp(-\theta S)]\big)^x dx\bigg) \\
&\leq \exp(\theta) \times 2^{\lceil p \rceil} \times \bigg(1 + \int_1^{\infty} x^{\lceil p \rceil} \big(\mathbb{E}[\exp(-\theta S)]\big)^x dx\bigg) \quad \text{since} \quad \frac{x+1}{x} \leq 2 \text{ for all } x \geq 1 \\
&\leq \exp(\theta) \times 2^{\lceil p \rceil} \times \bigg(1 + \int_0^{\infty} x^{\lceil p \rceil} \big(\mathbb{E}[\exp(-\theta S)]\big)^x dx\bigg) \\
&= \exp(\theta) \times 2^{\lceil p \rceil} \times \bigg(1 + \lceil p \rceil! \log^{-\big(\lceil p \rceil + 1\big)} \big(\frac{1}{\mathbb{E}[\exp(-\theta S)]}\big)\bigg) \\
&\leq \exp(\theta) \times 2^{\lceil p \rceil} \times \bigg(1 + \lceil p \rceil! \times \big(1 - \mathbb{E}[\exp(-\theta S)]\big)^{-(\lceil p \rceil + 1)}\bigg) \\
&\leq \exp(\theta) \big(\frac{24p}{1 - \mathbb{E}[\exp(-\theta S)]}\big)^{p+2},
\end{aligned}$$

with the second-to-last inequality following from the fact that $\log(\frac{1}{x}) \geq 1 - x$ for all $x \in (0, 1)$, and the final inequality follows from some straightforward algebra. Combining the above completes the proof. □

2.7.6 Proof of Corollary 2.8

We note that the proof of Corollary 2.8 follows nearly identically to the proof of Lemma 8 of [63], although with all quantities made explicit (and using the results of Lemma 2.12). We include the entire proof for completeness.

Proof of Corollary 2.8. Let S^e denote the first renewal interval in $\mathcal{N}_1$, and f_{S^e} its density function, whose existence is guaranteed by the basic properties of the equilibrium distribution. Observe that we may construct $\mathcal{N}_1$ and $\mathcal{N}_o$ on the same probability space so that for all $t \geq 0$, $N_1(t) = I(S^e \leq t) + N_o\big((t - S^e)^+\big)$, with $\mathcal{N}_o$ independent of S^e. Thus

$$N_1(t) - t = \Big(N_o\big((t - S^e)^+\big) - \big(t - S^e\big)^+\Big) + \Big(I\big(S^e \leq t\big) - \big(t - (t - S^e)^+\big)\Big).$$

Fixing some $t \geq 1$, it follows from (2.20) and the triangle inequality that $\mathbb{E}\big[|N_1(t) - t|^r\big]$ is at most

$$2^{r-1}\mathbb{E}\big[|N_o\big((t - S^e)^+\big) - \big(t - S^e\big)^+|^r\big] \tag{2.87}$$

$$+2^{r-1}\mathbb{E}\big[|I\big(S^e \leq t\big) - \big(t - (t - S^e)^+\big)|^r\big]. \tag{2.88}$$

We now bound the term $\mathbb{E}\big[|N_o\big((t-S^e)^+\big) - \big(t-S^e\big)^+|^r\big]$ appearing in (2.87), which equals

$$\int_0^{t-1} \mathbb{E}\big[|N_o\big(t - s\big) - \big(t - s\big)|^r\big] f_{S^e}(s)ds \tag{2.89}$$

$$+ \int_{t-1}^{t} \mathbb{E}\big[|N_o\big(t - s\big) - \big(t - s\big)|^r\big] f_{S^e}(s)ds. \tag{2.90}$$

Lemma 2.12 and Markov's inequality (after raising both sides to the rth power), combined

with our assumptions on r and t, implies that (2.89) is at most

$$\begin{aligned}
&\exp(\theta)\mathbb{E}[S^r]\big(\frac{10^5 r^2}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+1}\int_0^{t-1}(t-s)^{\frac{r}{2}}f_{S^e}(s)ds\\
\leq\ &\exp(\theta)\mathbb{E}[S^r]\big(\frac{10^5 r^2}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+1}t^{\frac{r}{2}}\int_0^{t-1}f_{S^e}(s)ds\\
\leq\ &\exp(\theta)\mathbb{E}[S^r]\big(\frac{10^5 r^2}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+1}t^{\frac{r}{2}}.
\end{aligned}$$

Since $t-s\leq 1$ implies w.p.1 $|N_o(t-s)-(t-s)|^r\leq|N_o(1)+1|^r$, it follows from (2.20) and Lemma 2.15 that the (2.90) is at most

$$\begin{aligned}
&\mathbb{E}\big[|N_o(1)+1|^r\big]\times\mathbb{P}\big(S^e\in[t-1,t]\big)\\
\leq\ &2^{r-1}\bigg(\mathbb{E}\big[\big(N_o(1)\big)^r\big]+1\bigg)\\
\leq\ &2^{r-1}\bigg(\exp(\theta)\big(\frac{24r}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+2}+1\bigg)\\
\leq\ &\exp(\theta)\big(\frac{48r}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+2}.
\end{aligned}$$

Combining our bounds for (2.89) and (2.90), with some straightforward algebra, we find that (2.87) is at most

$$\mathbb{E}[S^r]\exp(\theta)\big(\frac{2\times 10^5 r^2}{1-\mathbb{E}[\exp(-\theta S)]}\big)^{r+2}t^{\frac{r}{2}}. \tag{2.91}$$

We now bound (2.88), which by (2.20) is at most

$$\begin{aligned}
&2^{2r-2}\bigg(1+\mathbb{E}\big[|\big(t-(t-S^e)^+\big)|^r\big]\bigg)\\
\leq\ &2^{2r-2}\bigg(1+\big(\int_0^t s^r f_{S^e}(s)ds+\int_t^\infty t^r f_{S^e}(s)ds\big)\bigg).
\end{aligned} \tag{2.92}$$

It follows from the basic properties of the equilibrium distribution and Markov's inequality that for all $s\geq 0$,

$$f_{S^e}(s)=\mathbb{P}(S>s)\leq\mathbb{E}[S^r]s^{-r}.$$

Thus the term $\int_0^t s^r f_{S^e}(s)ds + \int_t^\infty t^r f_{S^e}(s)ds$ appearing in (2.92) is at most

$$\begin{aligned}\int_0^t s^r\big(\mathbb{E}[S^r]s^{-r}\big)ds + t^r\int_t^\infty \big(\mathbb{E}[S^r]s^{-r}\big)ds &= \mathbb{E}[S^r]\bigg(\int_0^t ds + t^r\int_t^\infty s^{-r}ds\bigg)\\ &= \mathbb{E}[S^r]\bigg(t + t^r(r-1)^{-1}t^{1-r}\bigg)\\ &\leq 2\mathbb{E}[S^r]t. \qquad (2.93)\end{aligned}$$

Using (2.91) to bound (2.87), and (2.93) and (2.92) to bound (2.88), and combining with some straightforward algebra completes the proof. □

CHAPTER 3

HEAVY-TAILED QUEUES IN THE HALFIN-WHITT REGIME

3.1 Introduction.

3.1.1 Halfin-Whitt regime and literature review

The staffing of large-scale queueing systems, and the associated trade-offs, are a fundamental problem in Operations Research. The insight that in many settings of interest one should scale the number of servers to exceed the arrival rate by a quantity on the order of the square-root of the arrival rate, i.e. the so-called square-root staffing rule, is by now well-known. This setting is formalized by the so-called Halfin-Whitt scaling regime for parallel server queueing systems, studied originally by Erlang [110] and Jagerman [111], and formally introduced by Halfin and Whitt [31], who studied the $GI/M/n$ system (for large n) when the traffic intensity ρ scales like $1 - Bn^{-\frac{1}{2}}$ for some strictly positive excess parameter B. There the authors prove weak convergence of the resulting queue-length process over compact time intervals, as well as weak convergence of the corresponding sequence of steady-state queue length distributions, when the queue-length of the nth system is normalized by $n^{\frac{1}{2}}$. Namely, in both the transient and steady-state regimes, the queue-length scales like $n^{\frac{1}{2}}$ in the Halfin-Whitt regime.

The original results of [31] have since been extended in many directions. Here we only review those results most relevant to our own investigations, and refer the interested reader to [83] for a comprehensive overview. The most general results in the transient regime are those of [61, 62], which (customized to the setting of our own investigations, i.e. single-class parallel multi-server queues with i.i.d. inter-arrivals and service times) prove that as long as the inter-arrival process satisfies a form of the central limit theorem on the scaling of $n^{\frac{1}{2}}$ (which will in general hold if the inter-arrival times have finite variance), and the

service time distribution has finite mean, then the associated sequence of queue-length processes, normalized by $n^{\frac{1}{2}}$, converges weakly to a non-trivial limiting process (if the system is initialized appropriately), described implicitly as the solution to a certain stochastic convolution equation.

As regards the scaling of the corresponding sequence of steady-state queue lengths, the most general known results are as follows. Assuming that inter-arrival times and service times have finite $2 + \epsilon$ moment for some $\epsilon > 0$, [63] proves that the associated sequence of steady-state queue-lengths, normalized by $n^{\frac{1}{2}}$, is tight. Under several additional technical assumptions, including that the service times have finite third moment, the very recent results of [112, 68] show that the associated sequence has a unique weak limit. Such a result was previously shown for the setting of service times with finite support in [46]. In the presence of Markovian abandonments, an analogous result has been proven for the case of phase-type service times. Indeed, in this setting [36] proved that the sequence of steady-state queue length distributions, normalized by $n^{\frac{1}{2}}$, is tight with an explicit weak limit which the authors characterize as an Ornstein-Uhlenbeck process with piece-wise linear drift. We note that although e.g. phase-type distributions are dense within the family of all distributions, due to the nature of the limits involved with the Halfin-Whitt regime, it is typically not clear how to translate results for such a restricted class of distributions to the general setting.

There has also been considerable interest in understanding the quality of Halfin-Whitt type approximations for finite n (as opposed to having results which only hold asymptotically). Such results include [38, 41, 37, 40, 39, 113]. We refer the reader to [113] for a detailed overview and discussion, and note that none of these results apply to the heavy-tailed setting. The very recent results of [113] provided the first simple and explicit bounds for multi-server queues that scale universally as $\frac{1}{1-\rho}$ across different notions of heavy traffic, including the Halfin-Whitt scaling. However, the main results of [113] assumed that both inter-arrival and service times have finite $2 + \epsilon$ moment for some $\epsilon > 0$.

3.1.2 Heavy tails in the Halfin-Whitt regime

A key insight from modern queueing theory is that when inter-arrival or service times have a heavy tail (i.e. the tail of the probability distribution does not decay exponentially), the underlying system behaves qualitatively different, e.g. it may exhibit long-range dependencies over time, and have a higher probability of rare events [58]. As many applications in modern service systems (e.g. length of stay in a hospital, length of time of a call) are potentially highly variable (e.g. due to prolongued illnesses, or having to resolve a complex IT problem), and may experience traffic which is bursty in nature (e.g. long periods of low activity followed by periods of high activity) [114], and several studies have empirically verified this phenomena in applications relevant to the Halfin-Whitt scaling [7, 59], it is important to understand how the presence of heavy tails changes the performance of parallel server queues in the Halfin-Whitt scaling regime. Although there is a vast literature on parallel server queues with heavy-tailed inter-arrival and/or service times (which we make no attempt to survey here, instead referring the reader to [60]), it seems that surprisingly, very little is known about how such systems behave qualitatively in the Halfin-Whitt regime.

We now survey what is known in this setting. The results of [61, 62] imply that when inter-arrival times have finite second moment (i.e. satisfy a classical central limit theorem) and service times have finite mean (but may have infinite $1+\epsilon$ moment for some $\epsilon \in (0,1)$), the associated sequence of transient queue-length processes, normalized by $n^{\frac{1}{2}}$, converges weakly (over compact time sets) to a non-trivial limiting process (if the system is initialized appropriately), described implicitly as the solution to a certain stochastic convolution equation.

[1] considers the case in which inter-arrival times have (asymptotically) a so-called pure Pareto tail with index $\alpha \in (1,2)$, i.e. $\lim_{x\to\infty} \frac{\mathbb{P}(A>x)}{x^{\alpha}} = C$ for some $\alpha \in (1,2)$ and $C \in (0,\infty)$, and service times are deterministic. In this case, [1] identifies a different scaling regime, a certain modification of the Halfin-Whitt scaling regime with the

scaling modified to account for the heavy-tailed inter-arrivals. In particular, Reed considers the associated sequence of $GI/D/n$ queues when the traffic intensity ρ scales like $1 - Bn^{-\frac{1}{\alpha}}$ for some strictly positive excess parameter B. In this case, Reed proves that the associated sequence of steady-state waiting time random variables, rescaled so as to be multiplied by $n^{1-\frac{1}{\alpha}}$, converges weakly to an explicit limiting distribution $\hat{W}$ characterized as the supremum of a certain infinite-horizon one-dimensional discrete-time random walk, i.e. a so-called α-stable random walk, with drift $-B$. Furthermore, although Reed does not explicitly prove it, it follows from an analysis nearly identical to that given in [115] that by the distributional Little's law (which is applicable since service times are deterministic), the sequence of steady-state queue-length distributions, normalized by $n^{\frac{1}{\alpha}}$, also converges to $\hat{W}$. Intuitively, the steady-state queue length in the nth system is thus approximately $\hat{W}n^{\frac{1}{\alpha}}$. Namely, for $\alpha < 2$, $n^{\frac{1}{2}}$ **is no longer the correct scaling**. This insight is quite interesting, although we note the important fact that Reed's results are restricted to the case of deterministic service times.

Essentially all other references in the literature to queues in the Halfin-Whitt regime with heavy tails are to open questions, which we now review. The question of tightness of the associated sequence of steady-state queue length distributions, normalized by $n^{\frac{1}{2}}$, is similarly left open when service times have infinite variance.

The explicit bounds of [113] for multi-server queues, which exhibit universal $\frac{1}{1-\rho}$ scaling across different heavy-traffic regimes (including the Halfin-Whitt scaling), left the extension to the heavy-tailed case open as well. However, we note that the results of [116, 117] prove that even for the single-server queue, $\frac{1}{1-\rho}$ **is no longer the correct scaling** as $\rho \uparrow 1$ when service times are heavy-tailed, where the correct scaling instead involves a different function of ρ depending on the tail of the service time distribution. Intriguingly, the transient results of [61, 62] show that in the Halfin-Whitt regime, even when service times are heavy-tailed, $\frac{1}{1-\rho}$ is the correct scaling (at least for the transient queue-length distribution), as in the Halfin-Whitt regime $\frac{1}{1-\rho}$ will scale as the square root of the number of

servers. As such, it seems that in the heavy-tailed setting, whether $\frac{1}{1-\rho}$ is the correct scaling depends heavily on precisely how one sends a sequence of queues into heavy traffic.

Indeed, it has been recognized in the literature that the order in which one takes limits plays a critical role in the heavy-tailed setting, e.g. when simultaneously looking at large-deviations behavior and heavy traffic, and such questions have been analyzed in [108] for the single-server setting. For the case of multiple servers, it is known that the interaction between the number of servers, the traffic intensity, and the large deviations behavior is very subtle [118]. Recently, several results have been proven as regards the large deviations behavior when the number of servers and traffic intensity are held fixed [119, 118, 120]. However, much less is known as regards how the large deviations behavior scales while simultaneously altering the number of servers and traffic intensity. Although some general explicit bounds are given in [57], building on the earlier work of [56], those bounds do not scale properly in the Halfin-Whitt scaling, and e.g. depend sensitively on certain parameters being non-integer (with the bounds degrading as those parameters approach integers). Several interesting bounds are also given in [121], which proves that in certain settings heavy-tailed service times lead to heavy-tailed waiting times. However, the upper bounds presented there do not scale correctly in the Halfin-Whitt regime (see e.g. [113] for a discussion of how bounds based on cyclic scheduling scale), while the implications of the proven lower bounds in the Halfin-Whitt regime are unclear. We also note that using a robust-optimization approach, a different family of bounds was developed for a non-stochastic model of multi-server queues with heavy tails in [91], although those bounds also do not scale appropriately in the Halfin-Whitt regime.

3.1.3 Questions for this work

The above discussion regarding heavy-tailed inter-arrival and service times in the Halfin-Whitt regime motivates the following questions.

Question 3.1. If the inter-arrival times have finite second moment but the service times only

have finite $1+\epsilon$ moment for some $\epsilon \in (0,1)$, is the sequence of steady-state queue lengths in the Halfin-Whitt regime, normalized by $n^{\frac{1}{2}}$, still tight? We note that a positive answer is known for the corresponding sequence of transient queue lengths (properly initialized) over a fixed compact time interval, but the corresponding question for the steady-state queues remains open.

Question 3.2. For the setting in which inter-arrival times have infinite variance, can the scaling regime described by Reed in [1], henceforth referred to as the Halfin-Whitt-Reed scaling regime, be extended from the setting of deterministic service times to the setting of general service time distributions? Do the same insights regarding tightness and asymptotic scaling hold?

Question 3.3. Is it possible to derive simple and explicit bounds for multi-server queues in the Halfin-Whitt regime, when service times are heavy-tailed? As all previous work on explicit, non-asymptotic bounds for queues in the Halfin-Whitt regime assumed service times have a finite second moment, these would be the first such explicit bounds in the heavy-tailed setting.

3.1.4 Main contribution

In this chapter, we provide positive answers to Questions 3.1 - 3.3.

Answer 3.1. We prove that, so long as inter-arrival times have finite second moment and service times have finite $1+\epsilon$ moment for some $\epsilon > 0$, the sequence of steady-state queue lengths, normalized by $n^{\frac{1}{2}}$, is tight. Namely, the presence of heavy-tailed service times does not interfere with the fact that the steady-state queue lengths scale like $n^{\frac{1}{2}}$.

Answer 3.2. We prove that the Halfin-Whitt-Reed regime can indeed be extended to the setting of generally distributed service times. In particular, we prove that when inter-arrival times have an asymptotically pure Pareto tail with index $\alpha \in (1,2)$, and procesesing times

have a finite $1+\epsilon$ moment for some $\epsilon > 0$, the sequence of steady-state queue lengths (under the Halfin-Whitt-Reed scaling), normalized by $n^{\frac{1}{\alpha}}$, is tight.

Answer 3.3. We extend the framework introduced in Chapter 2 ([113]) to provide the first simple and explicit bounds for multi-server queues that scale correctly in the Halfin-Whitt regime when service times are heavy-tailed.

3.1.5 Chapter outline

The rest of the chapter proceeds as follows. We state our main results in Section 3.2. We prove our explicit bounds for multi-server queues in the Halfin-Whitt regime when service times may be heavy-tailed in Section 3.3. We extend the analysis of Reed from the special case of deterministic service times to the case of general service times, i.e. generalizing the notion of the Halfin-Whitt-Reed regime, in Section 3.4. We provide a summary of our results and directions for future research in Section 3.5. Finally, we include a technical appendix in Section 3.6, which contains several technical arguments from throughout the chapter.

3.2 Main results

In this section we formally state our main results.

3.2.1 Additional Notations

We first introduce some additional notations for the Halfin-Whitt (-Reed) regime. In additional to the notations introduced in Section 1.5, let us fix $\mathbb{E}[A] = \mathbb{E}[S] = 1$ (recalling that A and S represent the inter-arrival time and service time respectively). For $B > 0, \alpha > 1$, and $n > B^{\frac{\alpha}{\alpha-1}}$, let $\lambda_{n,B,\alpha} \triangleq n - Bn^{\frac{1}{\alpha}}$, and $\mathcal{Q}^n_{A,S,B,\alpha}$ denote the FCFS $GI/GI/n$ queue with inter-arrival distribution $A\lambda^{-1}_{n,B,\alpha}$, and service time distribution S. If for any given initial condition, the total number of jobs in $\mathcal{Q}^n_{A,S,B,\alpha}$ (number in service + number waiting in queue) converges in distribution (as time goes to infinity, independent of the particular

initial condition) to a steady-state r.v. $Q^n_{A,S,B,\alpha}(\infty)$, we say that "$Q^n_{A,S,B,\alpha}(\infty)$ exists". For n large, $\mathcal{Q}^n_{A,S,B,\frac{1}{2}}$ is said to be in the Halfin-Whitt (a.k.a. Quality-and-efficiency driven, QED) scaling regime [31]. As Reed [1] had studied $\mathcal{Q}^n_{A,S,B,\alpha}$ for n large when S is deterministic and $\alpha \in (1,2)$, we will generally say that $\mathcal{Q}^n_{A,S,B,\alpha}$ is in the Halfin-Whitt-Reed regime when n is large and $\alpha \in (1,2)$. Similarly, if for any given initial condition, the waiting time (i.e. time in system between time of arrival and time at which service begins) for the jth job to arrive to $\mathcal{Q}^n_{A,S,B,\alpha}$ converges in distribution (as $j \to \infty$, independent of particular initial condition) to a steady-state r.v. $W^n_{A,S,B,\alpha}(\infty)$, we say that "$W^n_{A,S,B,\alpha}(\infty)$ exists".

Let us also introduce some notations for the stable law, which will be relevant when we describe the main results. More on stable law and the associated generalized central limit theorem is discussed in Section 3.4.1. Following the notations in [122, 29], let $S_\alpha(\sigma,\beta,\mu)$ denote a real-valued random variable having stable distribution, with stability parameter $\alpha > 0$, the scale parameter $\sigma > 0$, the skewness parameter $\beta \in [-1,1]$ and the shift parameter $\mu \in (-\infty,\infty)$. When $\alpha \in (1,2)$, $\mu = \mathbb{E}[S_\alpha(\sigma,\beta,\mu)] < \infty$ and $S_\alpha(\sigma,\beta,\mu)$ has characteristic function $\mathbb{E}[e^{i\theta S_\alpha(\sigma,\beta,\mu)}] = \exp(-(\sigma|\theta|)^\alpha(1 - i\beta(\text{sgn}\theta)\tan(\pi\alpha/2) + i\mu\theta))$, where sgn$\theta$ is the sign of $\theta \in \mathbb{R}$. Let $\hat{S}_\alpha(t,\beta,\mu)_{t\geq 0}$ denote an (α,β)-stable Levy motion s.t. $S_\alpha(0,\beta,\mu) = 0$ and $\hat{S}_\alpha(s+t,\beta,\mu) - \hat{S}_\alpha(s,\beta,\mu) \sim t^{\frac{1}{\alpha}} S_\alpha(1,\beta,\mu), \forall s,t \geq 0$.

Also, for two r.v.s X, Y, let $X \sim Y$ denote equivalence in distribution. In addition, for $t > 0$, let $\Gamma(t) \stackrel{\Delta}{=} \int_0^\infty x^{t-1}\exp(-x)dx$ denote the Γ-function. For $\alpha \in (1,2)$, let $C_\alpha \stackrel{\Delta}{=} (1-\alpha)\big(\Gamma(2-\alpha)\cos(\frac{\pi}{2}\alpha)\big)^{-1}$, where we note that $C_\alpha \in (0,\infty)$ for all $\alpha \in (1,2)$.

3.2.2 Main results

We begin by formalizing Answers 3.1 and 3.3, i.e. stating our simple and explicit bounds, as well as the implied tightness results. We note that our tightness results are essentially the best possible, as the results of [1] show that when inter-arrival times have infinite second moment square-root scaling is no longer appropriate. In particular, the only case left

unresolved is that in which $\mathbb{E}[S] < \infty$ but $\mathbb{E}[S^{1+\epsilon}] = \infty$ for all $\epsilon > 0$. Furthermore, even in that case, we believe our techniques could be extended to prove tightness and explicit bounds. For a real number x, let $x^+ \triangleq \max(x, 0)$.

Theorem 3.1 (Answer 3.3)**.** *Suppose that* $\mathbb{E}[A^2] < \infty$*, and* $\mathbb{E}[S^{1+\epsilon}] < \infty$ *for some* $\epsilon \in (0, 1]$*. Then for all* $B > 0$ *and* $n > 4B^2$ *such that* $Q^n_{A,S,B,2}(\infty)$ *exists, it holds that for all* $x \geq 16$*,* $\mathbb{P}\Big(n^{-\frac{1}{2}}\big(Q^n_{A,S,B,2}(\infty) - n\big) \geq x\Big)$ *is at most*

$$10^{100}\bigg(\epsilon\big(1 - \mathbb{E}[\exp(-S)]\big)\bigg)^{-7}(10\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{\epsilon}}(B^{-1} + B^{-2})x^{-\frac{\epsilon}{1+\epsilon}}.$$

Corollary 3.1 (Answer 3.1)**.** *Suppose that* $\mathbb{E}[A^2] < \infty$*,* $\mathbb{E}[S^{1+\epsilon}] < \infty$ *for some* $\epsilon \in (0, 1]$*, and for some* $B > 0$*,* $Q^n_{A,S,B,2}(\infty)$ *exists for all sufficiently large* n*. Then* $\big\{n^{-\frac{1}{2}}\big(Q^n_{A,S,B,2}(\infty) - n\big)^+, n > 4B^2\big\}$ *is tight.*

Finally, we formalize Answer 3.2, extending the Halfin-Whitt-Reed regime to generally distributed service times. First, we formalize the Halfin-Whitt-Reed scaling regime through an appropriate set of assumptions.

HWR-α Assumptions.

- $\alpha \in (1, 2)$*;*
- *There exists* $C_A \in (0, \infty)$ *s.t.* $\lim_{x\to\infty} x^\alpha \mathbb{P}(A > x) = C_A$*;*
- *There exists* $\epsilon \in (0, 1]$ *s.t.* $\mathbb{E}[S^{1+\epsilon}] < \infty$*;*
- *For each fixed* $B > 0$*,* $Q^n_{A,S,B,\alpha}(\infty)$ *and* $W^n_{A,S,B,\alpha}(\infty)$ *exists for all sufficiently large* n*.*

Then our formalization of Answer 3.2 is as follows.

Theorem 3.2 (Answer 3.2)**.** *If the HWR-α assumptions hold, then* $\left\{n^{-\frac{1}{\alpha}}Q^n_{A,S,B,\alpha}(\infty), n \geq 1\right\}$ *is tight. Furthermore, for all* $x > 0$*,*

$$\limsup_{n\to\infty} \mathbb{P}\bigg(n^{-\frac{1}{\alpha}}(Q^n_{A,S,B,\alpha}(\infty) - n) > x\bigg) \leq \mathbb{P}\bigg(\sup_{t\geq 0}\bigg(-(\frac{C_A}{C_\alpha})^{\frac{1}{\alpha}}\hat{S}_\alpha(t,1,0) - Bt\bigg) > x\bigg). \tag{3.1}$$

Note that our bound does not depend on the particulars of the service time distribution at all. As $-\hat{S}_\alpha(t,1,0)$ is a so-called spectrally negative Levy process (i.e. all jumps are negative), it is well-known that $\sup_{t\geq 0}\bigg(-(\frac{C_A}{C_\alpha})^{\frac{1}{\alpha}}\hat{S}_\alpha(t,1,0) - Bt\bigg)$ follows a simple exponential distribution (cf. [123]). In particular, we have the following corollary, which follows immediately from Theorem 3.2 and ([123]).

Corollary 3.2. *If the HWR-α assumptions hold, then for all* $x > 0$*,*

$$\limsup_{n\to\infty} \mathbb{P}\bigg(n^{-\frac{1}{\alpha}}(Q^n_{A,S,B,\alpha}(\infty) - n) > x\bigg) \leq \exp\big(-\big(\frac{B}{C_A\alpha\Gamma(-\alpha)}\big)^{\frac{1}{\alpha-1}}x\big). \tag{3.2}$$

We now proceed to prove the main results.

3.3 Explicit bounds and proof of Theorem 3.1

In this section we prove Theorem 3.1, from which our tightness result Corollary 3.1 will immediately follow. We proceed by extending the framework of [63, 113] to the heavy-tailed setting and will rely heavily on some of the bounds discussed in Chapter 2.

The first bound is already introduced in Theorem 2.2, which bounds the steady-state queue length of a $GI/GI/n$ by the supremum of a relatively simple one-dimensional random walk. Customized to our particular setting (i.e. in terms of the Halfin-Whitt(-Reed) regime), it states that

Theorem 3.3 (Theorem 2.2,[63])**.** *Suppose that* $B > 0, \alpha > 1$*,* $n > B^{\frac{\alpha}{\alpha-1}}$*, and* $Q^n_{A,S,B,\alpha}(\infty)$

exists. Then for all $x \geq 0$*,*

$$\mathbb{P}\bigg(n^{-\frac{1}{\alpha}}\big(Q^n_{A,S,B,\alpha}(\infty)-n\big)\geq x\bigg)\leq \mathbb{P}\bigg(n^{-\frac{1}{\alpha}}\sup_{t\geq 0}\bigg(A(\lambda_{n,B,\alpha}t)-\sum_{i=1}^{n}N_i(t)\bigg)\geq x\bigg). \quad (3.3)$$

We note that in Chapter 2, while the proven bounds for pooled renewal processes relied heavily on the assumption that $\mathbb{E}[S^2] < \infty$, some of the intermediate results such as Lemma 2.6 and Lemma 2.18 are also applicable even when the variance is infinite, which we will further exploit to provide the needed bounds even in the heavy-tailed setting.

3.3.1 Novel bound for variance of pooled heavy-tailed renewal processes

In this section, we prove a novel simple, explicit, and non-asymptotic bound for the variance of a heavy-tailed equilibrium renewal process, i.e. $Var[N_1(t)]$. We note that for the case $\mathbb{E}[S^2] < \infty$, both the renewal function (i.e. $\mathbb{E}[N_o(t)]$), and the variance of $N_1(t)$, are understood fairly precisely, with fairly tight bounds known (especially under further assumptions e.g. finite third moment, cf. [124, 125, 126, 83]). The correct asymptotic scaling is also known in the heavy-tailed setting, under additional assumptions such as that S is regularly varying, and/or belongs to the domain of attraction of an appropriate stable law (cf. [127, 128, 129, 130, 131]), and in some of our later large deviation results we will use certain of these precise asymptotics. We also note that the literature contains certain non-explicit general results regarding the central moments of $N_1(t)$ under minimal moment conditions, showing e.g. that $\mathbb{E}[S^{1+\epsilon}] < \infty$ implies that $\mathbb{E}[|\mathcal{N}_{o,1}(t) - \mu_S t|^{1+\epsilon}]$ is asymptotically sublinear in t (cf. [102]), although these results do not seem amenable to our analysis. Here we provide a different result (which is, to our knowledge, new) under minimal assumptions on S. The result builds on an elegant bounding argument of [132], and a well-known explicit integral representation for $Var[N_1(t)]$ (cf. [124, 125, 126, 29]).

Lemma 3.1. *Suppose that* $\mathbb{E}[S^{1+\epsilon}] < \infty$ *for some* $\epsilon \in (0, 1]$*. Then for all* $t \geq 0$*, it holds*

that

$$Var[N_1(t)] \leq (4\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{\epsilon}}\big(t + t^{1+\frac{1}{1+\epsilon}}\big).$$

Our proof proceeds by first expressing $Var[N_1(t)]$ in terms of an integral involving the renewal function, and then using a result of [132] to bound the renewal function (and the aforementioned integral). We begin by stating the desired integral representation.

Lemma 3.2 ([124, 125, 126, 29])**.** *For all $t \geq 0$, it holds that*

$$Var[N_1(t)] = 2\int_0^t \bigg(\big(\mathbb{E}[N_o(s)] + 1 - s\big) - \frac{1}{2}\bigg)ds.$$

We next state the appropriate result from [132], customized to our own setting. In particular, the following lemma follows immediately from [132] Theorem 2, by taking the function h defined there to be $h(x) = x^{1+\epsilon}$.

Lemma 3.3 ([132] Theorem 2)**.** *Suppose that $\mathbb{E}[S^{1+\epsilon}] < \infty$ for some $\epsilon \in (0, 1]$. Then for all $t \geq 0$, it holds that*

$$t - 1 \leq \mathbb{E}[N_o(t)] \leq t - 1 + (\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{1+\epsilon}}(\mathbb{E}[N_o(t)] + 1)^{\frac{1}{1+\epsilon}}. \tag{3.4}$$

We note that Lemma 3.3 does not directly provide an easily used bound for $\mathbb{E}[N_o(s)] + 1 - s$, as the right-hand-side of (3.4) is essentially a "recursive bound" for $\mathbb{E}[N_o(s)]$, i.e. $\mathbb{E}[N_o(s)]$ is bounded in terms of a different function of $\mathbb{E}[N_o(s)]$. We now show how to use Lemma 3.3 to provide explicit bounds for $\mathbb{E}[N_o(s)] + 1 - s$.

Corollary 3.3. *Under the same assumptions as Lemma 3.3, for all $t \geq 0$,*

$$\mathbb{E}[N_o(t)] + 1 - t \leq (2\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{\epsilon}}(1 + t^{\frac{1}{1+\epsilon}}).$$

Proof of Corollary 3.3. Let us fix $t \geq 0$. Letting $Y_t \stackrel{\Delta}{=} \mathbb{E}[N_o(t)] + 1 - t$, we conclude from

Lemma 3.3 that

$$0 \leq Y_t \leq (\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{1+\epsilon}}(Y_t + t)^{\frac{1}{1+\epsilon}}. \tag{3.5}$$

If $Y_t = 0$, we are done. Thus suppose $Y_t > 0$. Then (3.5) implies that

$$Y_t^{\frac{\epsilon}{1+\epsilon}} \leq (\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{1+\epsilon}}(1 + \frac{t}{Y_t})^{\frac{1}{1+\epsilon}}. \tag{3.6}$$

We first prove that $Y_t \leq \max\left(t, \left(2\mathbb{E}[S^{1+\epsilon}]\right)^{\frac{1}{\epsilon}}\right)$. Indeed, suppose for contradiction that $Y_t > \max\left(t, \left(2\mathbb{E}[S^{1+\epsilon}]\right)^{\frac{1}{\epsilon}}\right)$. Then (3.6) implies that

$$\left(2\mathbb{E}[S^{1+\epsilon}]\right)^{\frac{1}{1+\epsilon}} < (\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{1+\epsilon}} 2^{\frac{1}{1+\epsilon}},$$

itself a contradiction, thus proving the desired statement, which itself implies that

$$0 < Y_t \leq \left(2\mathbb{E}[S^{1+\epsilon}]\right)^{\frac{1}{\epsilon}} + t. \tag{3.7}$$

Plugging (3.7) into the right-hand-side of (3.5), applying the subadditivity of the function $f(x) = x^{\frac{1}{1+\epsilon}}$ (which follows from concavity), and the fact that $\mathbb{E}[S^{1+\epsilon}] \geq 1$ (by Jensen's inequality since $\mathbb{E}[S] = 1$), we find that

$$\begin{aligned}
Y_t &\leq (\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{1+\epsilon}}\left(\left(2\mathbb{E}[S^{1+\epsilon}]\right)^{\frac{1}{\epsilon}} + 2t\right)^{\frac{1}{1+\epsilon}} \\
&\leq (\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{1+\epsilon}}\left(\left(2\mathbb{E}[S^{1+\epsilon}]\right)^{\frac{1}{\epsilon}\times\frac{1}{1+\epsilon}} + (2t)^{\frac{1}{1+\epsilon}}\right) \\
&\leq 2^{\frac{1}{\epsilon}\times\frac{1}{1+\epsilon}} \times (\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{1+\epsilon}\times(1+\frac{1}{\epsilon})} \times (1 + t^{\frac{1}{1+\epsilon}}) \\
&\leq (2\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{\epsilon}}(1 + t^{\frac{1}{1+\epsilon}}),
\end{aligned}$$

completing the proof. □

With Lemma 3.2 and Corollary 3.3 in hand, we now complete the proof of Lemma 3.1.

Proof of Lemma 3.1. It follows from Lemma 3.2 and Corollary 3.3 that for all $t \geq 0$,

$$\begin{aligned} Var[N_1(t)] &\leq 2(2\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{\epsilon}} \int_0^t \left(1 + s^{\frac{1}{1+\epsilon}}\right) ds \\ &\leq (4\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{\epsilon}} \left(t + t^{1+\frac{1}{1+\epsilon}}\right), \end{aligned}$$

completing the proof. □

3.3.2 Proof of Theorem 3.1

In this section we complete the proof of Theorem 3.1. We first apply Theorem 3.3 by applying a straightforward union bound to the right-hand-side of (3.3), along with non-negativity and some basic monotonicities, to conclude the following.

Lemma 3.4. *Suppose that $B > 0, \alpha \in (1, 2]$, and $n > B^{\frac{\alpha}{\alpha-1}}$. Then for all $x \geq 0$, $\mathbb{P}\left(n^{-\frac{1}{\alpha}} \sup_{t \geq 0} \left(A(\lambda_{n,B,\alpha} t) - \sum_{i=1}^n N_i(t) \right) \geq x \right)$ is at most*

$$\mathbb{P}\left(n^{-\frac{1}{\alpha}} \sup_{t \geq 0} \left(A\big(\lambda_{n,B,\alpha} t\big) - (n - \frac{1}{2} B n^{\frac{1}{\alpha}}) t \right) \geq \frac{1}{2} x \right) \tag{3.8}$$

$$+ \quad \mathbb{P}\left(n^{-\frac{1}{2}} \sup_{t \geq 0} \left(\big(nt - \sum_{i=1}^n N_i(t)\big) - \frac{B}{2} n^{\frac{1}{2}} t \right) \geq \frac{1}{2} x \right). \tag{3.9}$$

Bounding (3.8), the supremum associated with the arrival process

We first bound (3.8). As here we want the most general result possible (i.e. only assuming finite second moment for the inter-arrival time distribution), we will proceed by relating the supremum to the waiting time in an appropriate single-server queue and applying Kingman's bound (as opposed to e.g. the analysis in [113] which required stronger moment assumptions). We begin with a simple observation, following from the basic properties of ordinary and equilibrium renewal processes. For $y > 1$, let W_y denote a r.v. distributed as the steady-state waiting time in a $GI/GI/1$ queue with inter-arrival times distributed as

yA and service times the constant 1.

Observation 1. *Suppose that $B > 0, \alpha \in (1, 2]$, $n > B^{\frac{\alpha}{\alpha-1}}$, and $Q^n_{A,S,B,\alpha}(\infty)$ exists. Then for all $\nu > \lambda_{n,B,\alpha}$ and $z \geq 0$,*

$$\mathbb{P}\left(n^{-\frac{1}{\alpha}} \sup_{t \geq 0} \left(A\big(\lambda_{n,B,\alpha} t\big) - \nu t \right) \geq z \right) \tag{3.10}$$

is at most

$$\mathbb{P}\left(n^{-\frac{1}{\alpha}} \sup_{k \geq 0} \big(k - \nu \sum_{i=1}^{k} \frac{A_i}{\lambda_{n,B,\alpha}} \big) \geq z - n^{-\frac{1}{\alpha}} \right).$$

It follows from Lindley's representation of the steady-state waiting time that (3.10) is at most

$$\mathbb{P}\left(n^{-\frac{1}{\alpha}} W_{\frac{\nu}{\lambda_{n,B,\alpha}}} \geq z - n^{-\frac{1}{\alpha}} \right).$$

Next, we recall the celebrated Kingman's bound for waiting times in a $GI/GI/1$ queue, only stating the result as customized to our particular setting.

Lemma 3.5 ([133], Kingman's Bound)**.** *Suppose that $\mathbb{E}[A^2] < \infty$. Then for all $y > 1$,*

$$\mathbb{E}[W_y] \leq \frac{y^2 \sigma_A^2}{2(y-1)}.$$

Combining Observation 1 (with $\nu = n - \frac{1}{2} B n^{\frac{1}{\alpha}}$), Lemma 3.5 (with $y = \frac{n - \frac{1}{2} B n^{\frac{1}{\alpha}}}{n - B n^{\frac{1}{\alpha}}}$), Markov's inequality, and some straightforward algebra (e.g. the fact that $x \geq 4$ implies $\frac{x}{2} - n^{-\frac{1}{\alpha}} \geq \frac{x}{4}$, and $n > (2B)^{\frac{\alpha}{\alpha-1}}$ implies $\frac{n - \frac{1}{2} B n^{\frac{1}{\alpha}}}{n - B n^{\frac{1}{\alpha}}} \leq 2$), we derive the following bound for (3.8).

Lemma 3.6. *Suppose that $\mathbb{E}[A^2] < \infty, B > 0, \alpha \in (1, 2]$, and $n > (2B)^{\frac{\alpha}{\alpha-1}}$. Then for all $x \geq 4$, (3.8) is at most*

$$10^2 \sigma_A^2 B^{-1} n^{1 - \frac{2}{\alpha}} x^{-1}.$$

Bounding (3.9), the supremum associated with the departure process

We proceed by using Lemma 3.1 to verify that the conditions of Lemma 2.6 hold for appropriate parameters, which we use to bound (3.9). In particular, we prove the following.

Lemma 3.7. *Suppose that* $\mathbb{E}[S^{1+\epsilon}] < \infty$ *for some* $\epsilon \in (0,1)$. *Then for all* $B > 0, n \geq 1$, *and* $x \geq 16$, *(3.9) is at most*

$$10^{92}\epsilon^{-7}(8\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{\epsilon}}\big(1-\mathbb{E}[\exp(-S)]\big)^{-5}(B^{-1}+B^{-2})x^{-\frac{\epsilon}{1+\epsilon}}.$$

Proof of Lemma 3.7. By Lemma 3.1, we find that for all $t \geq 1$,

$$\mathbb{E}\big[|\sum_{i=1}^{n} N_i(t) - nt|^2\big] \leq (8\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{\epsilon}} nt^{1+\frac{1}{1+\epsilon}}.$$

By Lemma 2.18, applied with $k = n, p = 3, \theta = 1$, we find that for all $t \in [0,1]$,

$$\mathbb{E}\big[|\sum_{i=1}^{n} N_i(t) - nt|^3\big] \leq \big(\frac{10^8}{1-\mathbb{E}[\exp(-S)]}\big)^5 \max\big(nt, (nt)^{\frac{3}{2}}\big).$$

Thus we find that the conditions of Lemma 2.6 are met with

$$C_1 = (8\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{\epsilon}} \quad , \quad C_2 = \big(\frac{10^8}{1-\mathbb{E}[\exp(-S)]}\big)^5 \quad , \quad r_1 = 2 \quad , \quad s_1 = 1+\frac{1}{1+\epsilon} \quad , \quad r_2 = 3.$$

Taking $\nu = \frac{B}{2}n^{\frac{1}{2}}$, $\lambda = \frac{x}{2}n^{\frac{1}{2}}$, we conclude that (3.9) is at most

$$\begin{aligned}&\Big(10^6\epsilon^{-1}\Big)^7 (8\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{\epsilon}}\big(\frac{10^8}{1-\mathbb{E}[\exp(-S)]}\big)^5 \\ &\times \quad \Big(n\big(\frac{B}{2}n^{\frac{1}{2}}\big)^{-(1+\frac{1}{1+\epsilon})}\big(\frac{x}{2}n^{\frac{1}{2}}\big)^{-\frac{\epsilon}{1+\epsilon}} + n^{\frac{3}{2}}\big(\frac{1}{4}xBn\big)^{-\frac{3}{2}}\Big).\end{aligned}$$

Combining with some straightforward algebra completes the proof. □

With Lemma 3.7 in hand, we now complete the proof of Theorem 3.1.

Proof of Theorem 3.1. Using Lemma 3.7 to bound (3.8), and Lemma 3.6 to bound (3.9), we conclude from Theorem 3.3 (after some straightforward algebra) that for all $B > 0$, $n > 4B^2$, and $x \geq 16$, $\mathbb{P}\Big(n^{-\frac{1}{2}}\big(Q^n_{A,S,B,2}(\infty) - n\big) \geq x\Big)$ is at most

$$10^{92}\epsilon^{-7}(8\mathbb{E}[S^{1+\epsilon}])^{\frac{1}{\epsilon}}\big(1-\mathbb{E}[\exp(-S)]\big)^{-5}(B^{-1}+B^{-2})x^{-\frac{\epsilon}{1+\epsilon}}$$
$$+\ 10^2\sigma_A^2 B^{-1}x^{-1}.$$

Combining with some straightforward algebra, and Theorem 3.3, completes the proof. □

3.4 The Halfin-Whitt-Reed regime, and proofs of Theorem 3.2 and Corollary 3.2

In this section, we generalize the analysis of Reed from [1] to the case of general service times, and call the corresponding scaling regime the Halfin-Whitt-Reed regime. First, we will need some additional background on so-called α-stable processes and the generalized central limit theorem.

3.4.1 α-stable processes and the generalized central limit theorem

The celebrated central limit theorem describes the behavior of normalized partial sums of i.i.d. random variables which have finite variance, and proves that the sequence of normalized sums converges in distribution to a standard normal r.v. In this section we describe the generalization of these results to the setting in which the variance is infinite. To do so, we must describe the family of variables to which such a sequence of normalized sums can converge, the family of so-called α-stable distributions. We note that many different parametrizations appear for these variables throughout the literature, and we stick to the conventions of [29]. An α-stable distribution is specified by four parameters: an index parameter $\alpha \in (0, 2]$, scale parameter $\sigma > 0$, skewness parameter $\beta \in [-1, 1]$, and location parameter $\mu \in (-\infty, \infty)$. We will let $S_\alpha(\sigma, \beta, \mu)$ denote a r.v. with the corresponding distribution. Then the generalized central limit theorem may be stated as follows. Here we

only state a special case which will suffice for our purposes, e.g. only treating the case involving a pure Pareto tail, only treating non-negative r.v.s, only treating the case $\alpha \in (1,2)$, etc. For $\alpha \in (1,2)$, let $C_\alpha \stackrel{\Delta}{=} (1-\alpha)\big(\Gamma(2-\alpha)\cos(\frac{\pi}{2}\alpha)\big)^{-1}$, where we note that $C_\alpha \in (0,\infty)$ for all $\alpha \in (1,2)$.

Theorem 3.4 (Generalized central limit theorem (cf. [29]))**.** *Suppose that* $\lim_{x\to\infty} x^\alpha \mathbb{P}(A > x) = C \in (0,\infty)$ *for some* $\alpha \in (1,2)$*. Then* $\{n^{-\frac{1}{\alpha}} \sum_{i=1}^n (A_i - 1), n \geq 1\}$ *converges in distribution to* $(\frac{C}{C_\alpha})^{\frac{1}{\alpha}} S_\alpha(1,1,0)$*, and we say that* A *belongs to the normal domain of attraction of this limiting r.v.*

There is also an analogous generalized version of the central limit theorem for renewal processes. Before stating this result, we will need to define the so-called α-stable Levy motion, a natural heavy-tailed generalization of Brownian motion (whose marginal distributions will be α-stable distributions). Recall that a Levy-process is a stochastic process with sample paths in the D-space (e.g. may have jumps) with stationary and independent increments, which takes value 0 at time 0, and has the property that for all $t \geq 0$, the probability that the process jumps at time t equals 0. A Levy-process $L(t)_{t\geq 0}$ is said to be an (α,β)-stable Levy motion with index parameter $\alpha \in (0,2]$ and skewness parameter $\beta \in [-1,1]$ if for all $t,s > 0$, $L(s+t) - L(s)$ is distributed as $t^{\frac{1}{\alpha}} S_\alpha(1,\beta,0)$. We denote this process by $\hat{S}_\alpha(t,\beta,0)$.

Theorem 3.5 (Generalized central limit theorem for renewal processes (cf. [29]))**.** *Under the same assumptions as Theorem 3.4,* $\big\{n^{-\frac{1}{\alpha}}\big(A_o(nt) - nt\big)_{0\leq t\leq T}, n \geq 1\big\}$ *and* $\big\{n^{-\frac{1}{\alpha}}\big(A(nt) - nt\big)_{0\leq t\leq T}, n \geq 1\big\}$ *both converge weakly, in the* M_1 *topology, to*

$$-(\frac{C}{C_\alpha})^{\frac{1}{\alpha}} \hat{S}_\alpha(t,1,0)_{0\leq t\leq T}.$$

3.4.2 Extending the Halfin-Whitt-Reed regime to general service times, and proof of Theorem 3.2

In this section we use our stochastic-comparison approach, and results associated with our explicit bounds (i.e. Theorem 3.1), to extend the Halfin-Whitt-Reed regime beyond the case of deterministic process times. In particular, we complete the proof of Theorem 3.2. We proceed by means of a series of lemmas, and begin by proving the needed tightness result.

Lemma 3.8. $\left\{n^{-\frac{1}{\alpha}} \sup_{t\geq 0}\left(A(\lambda_{n,B,\alpha}t) - \sum_{i=1}^{n} N_i(t)\right), n \geq 1\right\}$ *is tight.*

Proof of Lemma 3.8. By Lemma 3.4, it suffices to prove tightness (separately) of

$$\left\{n^{-\frac{1}{\alpha}} \sup_{t\geq 0}\left(A\big(\lambda_{n,B,\alpha}t\big) - (n - \frac{1}{2}Bn^{\frac{1}{\alpha}})t\right)\right\}, \tag{3.11}$$

and

$$\left\{n^{-\frac{1}{2}} \sup_{t\geq 0}\left(\big(nt - \sum_{i=1}^{n} N_i(t)\big) - \frac{B}{2}n^{\frac{1}{2}}t\right)\right\}. \tag{3.12}$$

As tightness of (3.12) follows immediately from Lemma 3.7, it suffices to demonstrate tightness of (3.11). However, tightness of (3.11) follows immediately from Observation 1, applied with $\nu = \frac{n-\frac{1}{2}Bn^{\frac{1}{\alpha}}}{n-Bn^{\frac{1}{\alpha}}}$, and Theorem 7.1 of [117], which gives sufficient conditions for tightness of the sequence of waiting times associated with a sequence of single-server queues with heavy-tailed inter-arrival times in heavy traffic. □

Next, we prove the appropriate weak convergence result.

Lemma 3.9. *For all* $T > 0$, $\left\{n^{-\frac{1}{\alpha}} \sup_{t\in[0,T]}\left(A(\lambda_{n,B,\alpha}t) - \sum_{i=1}^{n} N_i(t)\right), n \geq 1\right\}$ *converges weakly to* $\sup_{t\in[0,T]}\left(-(\frac{C}{C_\alpha})^{\frac{1}{\alpha}}\hat{S}_\alpha(t,1,0) - Bt\right)$.

Proof of Lemma 3.9. Note that $n^{-\frac{1}{\alpha}}\Big(A(\lambda_{n,B,\alpha}t) - \sum_{i=1}^{n} N_i(t)\Big)$ equals

$$\begin{aligned}
&(\frac{\lambda_{n,B,\alpha}}{n})^{\frac{1}{\alpha}} \times \lambda_{n,B,\alpha}^{-\frac{1}{\alpha}}\Big(A(\lambda_{n,B,\alpha}t) - \lambda_{n,b,\alpha}t\Big) \\
+ \quad & n^{\frac{1}{2}-\frac{1}{\alpha}} \times n^{-\frac{1}{2}}\big(nt - \sum_{i=1}^{n} N_i(t)\big) \\
- \quad & Bt.
\end{aligned}$$

Combining with Theorem 3.5, Lemma 3.7, and the basic properties of J_1 and M_1 convergence, e.g. continuity of the supremum map (cf. [29]), completes the proof of the desired weak convergence. □

To prove Theorem 3.2 we want to extend Lemma 3.9 to show the convergence of the all time supremum. Note that such a result is not immediate, as the framework of weak convergence (of stochastic processes) generally deals only with compact time intervals, so extra care must be taken to handle such an infinite time horizon. We also note that closely related ideas were used in the proof of Lemma 7 and Theorem 2 in [63], although their proofs made use of the service time distribution having finite second moment, and our result is stated in considerably greater generality. Now, we prove the following general result, giving sufficient conditions under which the convergence of the supremum of some (stationary) process over compact interval can be extended to the supremum over the entire real line. We defer all relevant proofs to the appendix.

Lemma 3.10. *Suppose that $\{Y_n(t)_{t\geq 0}, n \geq 1\}$ is a sequence of stochastic processes on $D[0,\infty)$ with stationary increments, and that $\mathcal{Y}_\infty(t)_{t\geq 0}$ is a fixed stochastic process (also with stationary increments, on $D[0,\infty)$). Suppose also that:*

1. *$Y_n(0) = 0$ w.p.1 for all $n \geq 1$;*
2. *$\{\sup_{t\geq 0} Y_n(t), n \geq 1\}$ is tight;*
3. *For all $M > 0$, $\lim_{t\to\infty} \mathbb{P}\big(\mathcal{Y}_\infty(t) \geq -M\big) = 0$;*

4. *For each fixed* $T > 0$, $\{\sup_{0\le t\le T} Y_n(t), n \ge 1\}$ *converges weakly to* $\sup_{0\le t\le T} \mathcal{Y}_\infty(t)$.

Then $\{\sup_{t\ge 0} Y_n(t), n \ge 1\}$ *converges weakly to* $\sup_{t\ge 0} \mathcal{Y}_\infty(t)$.

With Lemmas 3.8, 3.9 and 3.10 in hand, we now complete the proof of Theorem 3.2.

Proof of Theorem 3.2. In light of Theorem 3.3, it suffices to verify that $\left\{ n^{-\frac{1}{\alpha}} \left(A(\lambda_{n,B,\alpha} t) - \sum_{i=1}^{n} N_i(t) \right)_{t\ge 0}, n \ge 1 \right\}$ satisfies the conditions of Lemma 3.10. In light of Lemmas 3.8 and 3.9, it suffices to verify condition 3. It follows from the basic properties of Levy processes that $\hat{S}_\alpha(t, 1, 0)$ has the same distribution as $t^{\frac{1}{\alpha}} S_\alpha(1, 1, 0)$, and thus for all $M > 0$ and $t > 0$, $\mathbb{P}\big(\mathcal{Y}_\infty(t) \ge -M\big)$ equals

$$\mathbb{P}\bigg(-(\frac{C}{C_\alpha})^{\frac{1}{\alpha}} t^{\frac{1}{\alpha}} S_\alpha(1, 1, 0) \ge M + Bt \bigg).$$

Condition 3 then follows from the fact that $\alpha > 1$, and $S_\alpha(1, 1, 0)$ is a.s. finite. Combining the above verifies that the conditions of Lemma 3.10 are met, completing the proof. □

With Theorem 3.2 established, we now prove Corollary 3.2 using a well-known result that the all time supremum of a spectrally negative Levy process follows a simple exponential distribution.

Proof of Corollary 3.2. Let $X(t) \equiv -(\frac{C}{C_\alpha})^{\frac{1}{\alpha}} \hat{S}_\alpha(t, 1, 0) - Bt$. It suffices to show $\sup_{t\ge 0} X(t) \sim expo((\frac{B}{C_A \alpha \Gamma(-\alpha)})^{\frac{1}{\alpha-1}})$. As the process $\{X(t), t \ge 0\}$ is indeed a negative-drifted Levy stable motion with no positive jumps (cf. [29]), [123, Proposition 5] states that $\sup_{t\ge 0} X(t)$ distributed exponentially. Now we find the exact rate. Recalling when $\alpha \in (1, 2)$, the characteristic function for $S_\alpha(1, 1, 0)$ is

$$\varphi(s) = \mathbb{E}[e^{isS_\alpha(1,1,0)}] = \exp(-|s|^\alpha (1 + i \,\mathrm{sgn}(s) \tan(\frac{\pi\alpha}{2})),$$

we have

$$\mathbb{E}[e^{isX(t)}] = \mathbb{E}[e^{is(-\frac{C_A}{C_\alpha}t^{\frac{1}{\alpha}}S_\alpha(1,1,0)-Bt)}] = \varphi(-s\frac{C_A}{C_\alpha}t^{\frac{1}{\alpha}})e^{-isBt},$$

which for $s \geq 0$ and $t \geq 0$, can be simplified as (cf. [123, Section 8])

$$\mathbb{E}[e^{isX(t)}] = e^{t(K_\alpha(si)^\alpha)-iBs}, \;\; with \;\; K_\alpha \triangleq -(C_A/C_\alpha)\sec(\frac{\pi\alpha}{2}).$$

Then for $s \geq 0, t \geq 0$, we have the analytic continuation $\mathbb{E}[e^{sX(t)}] = e^{t(K_\alpha s^\alpha - Bs)} = e^{t\psi(s)}$, when setting $\psi(s) \triangleq K_\alpha s^\alpha - Bs$. It then follows from [123, Proposition 5] that $\sup_{t\geq 0} X(t) \sim expo(\gamma)$, where γ is the largest (positive) root solving $\psi(s) = K_\alpha s^\alpha - Bs = 0$, namely, $\gamma = (\frac{B}{K_\alpha})^{\frac{1}{\alpha-1}}$. It then follows from the definition of C_α and the fact that $\Gamma(s) = (s-1)\Gamma(s-1)$ for all complex number s that is not integers less than or equal to zero, that $\gamma = (\frac{B}{C_A\alpha\Gamma(-\alpha)})^{\frac{1}{\alpha-1}}$, concluding the proof. □

3.5 Conclusion

In this chapter, we provided the first analysis of steady-state multi-server queues in the Halfin-Whitt regime when service times have infinite variance. We proved that under minimal assumptions, i.e. only that service times have finite $1+\epsilon$ moment for some $\epsilon > 0$ and inter-arrival times have finite second moment, the sequence of stationary queue length distributions, normalized by $n^{\frac{1}{2}}$, is tight in the Halfin-Whitt regime. This confirmed that the presence of heavy tails in the service time distributions does not change the fundamental scaling of the steady-state queue length, as $n^{\frac{1}{2}}$ was also the correct scaling in the light-tailed case. Furthermore, we developed simple and explicit bounds for the steady-state queue in multi-server queues in the Halfin-Whitt regime, under only these minimal assumptions. Also, for the setting where instead the inter-arrival times have an asymptotically Pareto tail with index $\alpha \in (1,2)$, we extended recent results of [1] (who analyzed the case of deterministic service times) by proving that for general service time distributions, the sequence of stationary queue length distributions, normalized by $n^{\frac{1}{\alpha}}$, is tight (here we used

the scaling of [1], which we named the Halfin-Whitt-Reed scaling regime). Interestingly, our derived bounds do not depend at all on the specifics of the service time distribution, and are nearly tight even for the case of deterministic service times.

Our work leaves several interesting directions for future research. Even within the Halfin-Whitt regime, there is the obvious question of deriving tighter explicit bounds, e.g. doing away with the massive constant appearing in our bounds, and developing tighter bounds on the demonstrated tail decay rate. Even more interesting is the question of deriving any kind of simple and explicit bounds that scale universally across different notions of heavy traffic, as was accomplished under the assumption of a finite $2 + \epsilon$ moment in [113]. As the heavy-tailed service times can cause very different types of scaling across different notions of heavy traffic, deriving such bounds would likely require fundamentally different methods of analysis. Another question along these lines is to develop a clearer understanding of the connection (under e.g. the Halfin-Whitt scaling) between the finiteness of moments of the steady-state queue length, and how those moments scale with the traffic intensity. Although the question of which moments are finite is by now fairly well understood [56], the question of how those finite moments scale in heavy traffic remains largely open, where we note that some interesting progress there follows from the recent results of [113].

3.6 Appendix

3.6.1 Proof of Lemma 3.10

Proof of Lemma 3.10. First, we claim that

$$\lim_{T\to\infty} \limsup_{n\to\infty} \mathbb{P}\big(\sup_{t\geq T} Y_n(t) \geq 0\big) = 0. \tag{3.13}$$

Indeed, for all $M > 0$, by a union bound and stationary increments $\mathbb{P}\big(\sup_{t\geq T} Y_n(t) \geq 0\big)$ is at most

$$\mathbb{P}\big(Y_n(T) \geq -M\big) + \mathbb{P}\big(\sup_{t\geq 0} Y_n(t) \geq M\big). \tag{3.14}$$

It follows from (2) - (3) that for any given $\epsilon > 0$, we may select $M_\epsilon, T_\epsilon \in (0,\infty)$ s.t. $\mathbb{P}\big(\mathcal{Y}_\infty(T_\epsilon) \geq -M_\epsilon\big) < \frac{\epsilon}{2}$, $\limsup_{n\to\infty} \mathbb{P}\big(\sup_{t\geq 0} Y_n(t) \geq M_\epsilon\big) < \frac{\epsilon}{2}$. Combining with (4), (3.14), and the monotonicity of the supremum operator, it follows that for all $T \geq T_\epsilon$, $\limsup_{n\to\infty} \mathbb{P}\big(\sup_{t\geq T} Y_n(t) \geq 0\big) < \epsilon$. Combining with the definition of limit completes the proof of (3.13).

It follows from (3.13) that for any $x \geq 0$, we may construct a strictly increasing sequence of integers $\{T_{x,k^{-1}}, k \geq 1\}$ s.t. for all $k \geq 1$,

$$\limsup_{n\to\infty} \mathbb{P}\bigg(\sup_{t\geq T_{x,k^{-1}}} Y_n(t) \geq x\bigg) < k^{-1}.$$

Thus by a union bound, for all $x \geq 0$ and $k \geq 1$,

$$\limsup_{n\to\infty} \mathbb{P}\bigg(\sup_{t\geq 0} Y_n(t) \geq x\bigg) \leq \limsup_{n\to\infty} \mathbb{P}\bigg(\sup_{0\leq t\leq T_{x,k^{-1}}} Y_n(t) \geq x\bigg) + k^{-1}.$$

By letting $k \to \infty$, and applying (4), the monotonicity of the supremum operator, and the Portmanteau Theorem, we conclude that for all $x \geq 0$,

$$\limsup_{n\to\infty} \mathbb{P}\bigg(\sup_{t\geq 0} Y_n(t) \geq x\bigg) \leq \mathbb{P}\big(\sup_{t\geq 0} \mathcal{Y}_\infty(t) \geq x\big). \tag{3.15}$$

Next, we prove the analagous result for $\liminf_{n\to\infty} \mathbb{P}\bigg(\sup_{t\geq 0} Y_n(t) > x\bigg)$. In particular, for any fixed T, (4), the monotonicity of the supremum operator, and the Portmanteau

Theorem imply that

$$\liminf_{n\to\infty} \mathbb{P}\Big(\sup_{t\geq 0} Y_n(t) > x \Big) \geq \mathbb{P}\big(\sup_{t\in[0,T]} \mathcal{Y}_\infty(t) > x \big).$$

Combining with the monotonicity of the supremum operator, and letting $T \to \infty$, it follows that for all $x \geq 0$,

$$\liminf_{n\to\infty} \mathbb{P}\Big(\sup_{t\geq 0} Y_n(t) > x \Big) \geq \mathbb{P}\big(\sup_{t\geq 0} \mathcal{Y}_\infty(t) > x \big). \tag{3.16}$$

Combining (3.15) and (3.16), with the definition of weak convergence, completes the proof.

□

CHAPTER 4

LARGE DEVIATIONS FOR HEAVY-TAILED QUEUES IN THE HALFIN-WHITT REGIME

4.1 Introduction.

4.1.1 Large deviations in Halfin-Whitt regime

To study the likelihood of rare events and develop metrics to quantify their consequences are among the most fundamental tasks when designing and analyzing a stochastic model, as those events, however unlikely, may have serious consequences. Large deviation theory provides ways to study those rare events, generally through analyzing the asymptotic behavior of the tail probabilities of sequences of random variables. Although vast literature have been developed to provide theoretical probabilities to general analysis of the large deviation behavior (e.g. [134, 135, 136]), explicit solutions to any particular stochastic model often remain difficult to get. Large deviation is a potent tool for analyzing queueing system in steady state (e.g. [137, 138, 139, 140]), not only because it is of practical interest to understand some key performance metrics (e.g. stationary waiting time and queue length) under extreme situations, it is also because sometimes (bounds on) the tail behaviors of those performance metrics (large deviation behaviors) are relatively easier to study, if not the only ones analytically available.

Under the Halfin-Whitt regime, outside of the case of exponentially distributed service times, the known characterizations for the limiting process (when such a limit is known to exist) are quite complicated. As such, considerable effort has gone into understanding certain properties of this limit, where many of these results have pertained to the large deviations behavior of the limiting process. In particular, for the case of inter-arrival times with finite second moment and service times with finite support, [46] prove that the weak limit

(associated with the sequence of normalized steady-state queue lengths) has an exponential tail, with a precise exponent identified as $-\frac{2B}{c_A^2+c_S^2}$, where $c_A^2(c_S^2)$ is the squared coefficient of variation (s.c.v) of inter-arrival (service) times. Namely, they prove that under those assumptions, the associated weak limit $\hat{Q}$ satisfies $\lim_{x\to\infty} x^{-1}\log\left(\mathbb{P}(\hat{Q}>x)\right) = -\frac{2B}{c_A^2+c_S^2}$. Put another way, the probability that the limiting process exceeds a large value x behaves (roughly up to exponential order) like $\exp\big(-\frac{2B}{c_A^2+c_S^2}x\big)$. The known results for the case of exponentially distributed and H_2^* service times yields the same exponent. The stochastic comparison approach of [63] was able to prove that the same exponent yields an upper bound on the large deviations behavior of any subsequential limit of the associated sequence of normalized queue-length random variables assuming only that there exists $\epsilon > 0$ s.t. inter-arrival and service times have finite $2+\epsilon$ moments, with equality for the case of exponentially distributed inter-arrival times. Far less is known when it comes to the large deviation behavior associated with the queueing systems with heavy-tailed inter-arrival or/and service time distribution under Halfin-Whitt regime.

In [63], the authors note that the identified limiting large deviations exponent $-\frac{2B}{c_A^2+c_S^2}$ equals zero when either inter-arrival or service times have infinite variance, and leave as an open question identifying the correct behavior in the presence of heavy tails. **We further investigate this open question in the chapter**. This is particularly interesting to investigate as the vanishing large deviation exponent suggests that a fundamentally different behavior may arise for the (properly scaled) steady-state queue length in the presence of heavy tails, a subject that was previously cast light on in [121], where the steady-state waiting time of the $M/GI/s$ queue was studied, and it was shown that the steady-state waiting time distribution "inherits" the heavy tails of the service time distribution.

To be more specific, in this chapter, complementing the tightness results established in Chapter 3, we will investigate the large deviation behavior of the scaled limiting process under $1)$ traditional Halfin-Whitt regime when service times are heavy-tailed (Pareto tail with index $\alpha \in (1,2)$) and inter-arrival times have finite second moment and $2)$ under the

Halfin-Whitt-Reed regime when inter-arrival times are heavy-tailed (Pareto tail with index $\alpha \in (1,2)$) and service times have at least $1+\epsilon$ moment for some $\epsilon \in (0,1]$. Note that the tightness of the sequences associated to above two cases were proven in Corollary 3.1 and Theorem 3.2 respectively.

4.1.2 Main contribution

First, for the special case that the inter-arrival times have finite second moment and the service times have an asymptotically pure Pareto tail, i.e. $\lim_{x\to\infty} \frac{\mathbb{P}(S>x)}{x^\alpha} = C$ for some $\alpha \in (1,2)$ and $C \in (0,\infty)$, we explicitly bound the large deviations behavior of the corresponding diffusion limit (under $n^{\frac{1}{2}}$ scaling, as was established in Corollary 3.1). In particular, we prove that the tail has a *subexponential decay*, i.e. that $\lim_{x\to\infty} x^{1-\alpha} \log\left(\mathbb{P}\big(\hat{Q}>x\big)\right)$ is at most an explicit strictly negative constant. Furthermore, for the case of Markovian inter-arrival times, we prove a lower bound which certifies that this is indeed the exact large deviations behavior. Interestingly, in contrast to the light-tailed (i.e. finite variance) setting, here we find that rare events are fundamentally more likely, with the probability of seeing a large queue length $xn^{\frac{1}{2}}$ decaying like $\exp(-C'x^{\alpha-1})$ with $\alpha-1 \in (0,1)$ and C' an explicit constant. This in essence resolves the question of the previously identified large deviations exponent $-\frac{2B}{c_A^2+c_S^2}$ which vanishes in the infinite-variance setting, since $\lim_{x\to\infty} \frac{C'x^{\alpha-1}}{x} = 0$. From a practical standpoint, this insight is important, as it suggests that when service times are heavy-tailed (which as noted is a setting relevant in several service-system applications), it is much more likely to see large queue lengths, where we successfully quantify the meaning of "much more likely".

Second, we further investigate the quality of the bound in Corollary 3.2. We show that when inter-arrival times have an asymptotically pure Pareto tail for some $\alpha \in (1,2)$ and service times have at least $1+\epsilon$ moment for some $\epsilon \in (0,1]$, under Halfin-Whitt-Reed regime, the upper bound stating that the right tail of the limiting distribution of steady state queue length (under $n^{\frac{1}{\alpha}}$ scaling) decays at least exponentially fast with rate being

some explicit constant, i.e. that $\lim_{x\to\infty} x^{-1}\log\Big(\mathbb{P}(\hat{Q}>x)\Big)$ is at most an explicit strictly negative constant, is in some sense nearly tight. Indeed, for the case of deterministic service times, we prove a lower bound that exactly matches the upper bound.

4.1.3 Chapter outline

The rest of the chapter proceeds as follows. We state our main results in Section 4.2. We prove our large deviations bounds for the setting that inter-arrival times have finite variance and service times are heavy-tailed in Section 4.3. The large deviations bounds for the setting that inter-arrival times are heavy-tailed under the generalized Halfin-Whitt-Reed regime are studied in Section 4.4. We provide a summary of our results and directions for future research in Section 4.5. Finally, we include a technical appendix in Section 4.6.

4.2 Main results

In this section, we formally state our main results. All notations are inherited from Section 3.2.1 (in addition to Section 1.5).

We begin by formulating a particular set of assumptions which we will need to state our results (which should be taken in addition to any assumptions posited to hold throughout the entire chapter, e.g. $\mathbb{E}[A]=\mathbb{E}[S]=1$).

GH1 Assumptions.

- $\mathbb{E}[A^2]<\infty$*;*
- *There exists* $\alpha_S\in(1,2)$ *and* $C_S\in(0,\infty)$ *s.t.* $\lim_{x\to\infty} x^{\alpha_S}\mathbb{P}(S>x)=C_S$*;*
- $\limsup_{t\downarrow 0} t^{-1}\big(\mathbb{P}(S\le t)-\mathbb{P}(S=0)\big)<\infty$*;*
- *For each fixed* $B>0$*,* $Q^n_{A,S,B,2}(\infty)$ *exists for all sufficiently large* n*.*

Let

$$C_{B,S}\stackrel{\Delta}{=}-C_S^{-1}B^{3-\alpha_S}(\frac{\alpha_S-1}{3-\alpha_S})^{2-\alpha_S}(2-\alpha_S).$$

Then our large deviation results when service times are asymptotically Pareto and inter-arrival times have finite second moment may be formalized as follows.

Theorem 4.1. *Under the GH1 Assumptions,*

$$\limsup_{x\to\infty} x^{-(\alpha_S-1)} \log\left(\limsup_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{2}}(Q^n_{A,S,B,2}(\infty)-n) > x\right)\right) \leq C_{B,S}.$$

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

$$\begin{aligned} &\liminf_{x\to\infty} x^{-(\alpha_S-1)} \log\left(\liminf_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{2}}(Q^n_{A,S,B,2}(\infty)-n) > x\right)\right) \\ = &\limsup_{x\to\infty} x^{-(\alpha_S-1)} \log\left(\limsup_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{2}}(Q^n_{A,S,B,2}(\infty)-n) > x\right)\right) \quad = \quad C_{B,S}. \end{aligned}$$

Roughly, Theorem 4.1 implies that when service times are asymptotically Pareto with power law decay parameter $\alpha_S \in (1,2)$, the probability of the queue exceeding a large queue length $xn^{\frac{1}{2}}$ decays roughly as $\exp\left(-C_{B,S}x^{\alpha_S-1}\right)$, which (since $\alpha_S - 1 < 1$) decays sub-exponentially. Namely, rare events are much more likely in this setting, as opposed to the light-tailed setting analyzed in [63], for which the decay was exponential. Note that $C_{B,S}$ is increasing in B, and hence in some sense seeing large queue lengths become "less likely" as B increases, which makes sense as when B is large the system is less loaded, where we note that a similar monotonicity was observed in [63]. Interestingly, the variability of the inter-arrival times does not appear in $C_{B,S}$, in contrast to the exponent identified in [63] for the light-tailed setting.

Recall

HWR-α Assumptions.

- $\alpha \in (1,2)$*;*
- *There exists* $C_A \in (0,\infty)$ *s.t.* $\lim_{x\to\infty} x^{\alpha}\mathbb{P}(A > x) = C_A$*;*
- *There exists* $\epsilon \in (0,1]$ *s.t.* $\mathbb{E}[S^{1+\epsilon}] < \infty$*;*

- *For each fixed $B > 0$, $Q^n_{A,S,B,\alpha}(\infty)$ and $W^n_{A,S,B,\alpha}(\infty)$ exists for all sufficiently large n.*

Then our large deviation results when inter-arrival times are asymptotically Pareto and service times have at least $1+\epsilon$ moment for some $\epsilon \in (0,1]$ may be formalized as follows.

Theorem 4.2. *Under the HWR-α Assumptions,*

$$\limsup_{x\to\infty} x^{-1} \log\left(\limsup_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{\alpha}}(Q^n_{A,S,B,\alpha}(\infty) - n) > x\right)\right) \leq -\left(\frac{B}{C_A \alpha \Gamma(-\alpha)}\right)^{\frac{1}{\alpha-1}}.$$

If in addition S is deterministic, namely the system is $GI/D/n$, then

$$\begin{aligned} &\liminf_{x\to\infty} x^{-1} \log\left(\liminf_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{\alpha}}(Q^n_{A,S,B,\alpha}(\infty) - n) > x\right)\right) \\ = &\limsup_{x\to\infty} x^{-1} \log\left(\limsup_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{\alpha}}(Q^n_{A,S,B,\alpha}(\infty) - n) > x\right)\right) \quad = \quad -\left(\frac{B}{C_A \alpha \Gamma(-\alpha)}\right)^{\frac{1}{\alpha-1}}. \end{aligned}$$

Theorem 4.2 shows that the bound on the large deviation exponent is in some sense tight, with an exact match in the case of deterministic service times. Note that the upper bound on the large deviation exponent, $-\left(\frac{B}{C_A \alpha \Gamma(-\alpha)}\right)^{\frac{1}{\alpha-1}}$, has nothing to do with the service time distribution. As a matter of fact, during the proof of Theorem 3.2 where the tightness and weak limit of $\left\{n^{-\frac{1}{\alpha}}(Q^n_{A,S,B,\alpha}(\infty) - n), n \geq 1\right\}$ were established, the service time distribution played virtually no roles but washed away. Thus it is logical to wonder if the bound is tight for more general classes of service time distributions. We leave it as an open question to explore in the future.

4.3 Large deviations under GH1 Assumptions, and proof of Theorem 4.1

In this section, we prove our large deviations results for the setting in which $\mathbb{E}[A^2] < \infty$ and S is asymptotically Pareto with infinite variance, i.e. Theorem 4.1. Our proof proceeds in a manner analogous to the large deviations results proven in [63]. Recall the stochastic

comparison upper bound (2.4, 3.3) and write it in the Halfin-Whitt-Reed regime notation (with $\alpha = 2$), we have

$$\mathbb{P}\bigg(n^{-\frac{1}{\alpha}}\big(Q^n_{A,S,B,2}(\infty) - n\big) \geq x\bigg) \leq \mathbb{P}\bigg(n^{-\frac{1}{\alpha}} \sup_{t\geq 0}\bigg(A(\lambda_{n,B,2}t) - \sum_{i=1}^{n} N_i(t)\bigg) \geq x\bigg). \quad (4.1)$$

In particular, we will use our tightness result (Corollary 3.1) to prove that our bounds for $Q^n_{A,S,B,2}(\infty) - n$ (4.1) behave like certain Gaussian processes in the Halfin-Whitt regime, where we note that (as in [63]) some care will have to be taken as these bounds are in the form of suprema over an infinite time horizon. We will then use known results from the theory of Gaussian processes and heavy-tailed renewal processes to derive the appropriate large deviations behavior.

4.3.1 Preliminary weak convergence results.

Before embarking on the proof of Theorem 4.1, we establish some preliminary weak convergence results to aid in our analysis. For an excellent review of weak convergence, and the associated spaces (e.g. $D[0,T]$) and topologies/metrics (e.g. uniform, J_1), we refer the reader to [29]. 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 $Z(t)_{t\geq 0}$ is known to be uniquely determined by its mean function $\mathbb{E}[Z(t)]_{t\geq 0}$ and covariance function $\mathbb{E}[Z(s)Z(t)]_{s,t\geq 0}$, and refer the reader to [141], and the references therein for details on existence, continuity, etc. Let $\aleph(t)_{t\geq 0}$ denote the w.p.1 continuous Gaussian process s.t. $\mathbb{E}[\aleph(t)] = 0, \mathbb{E}[\aleph(s)\aleph(t)] = c_A^2 \min(s,t)$, namely a driftless Brownian motion. Then we may conclude the following from the well-known Functional Central Limit Theorem (FCLT) for renewal processes (see [29] Theorem 4.3.2 and Corollary 7.3.1)

Theorem 4.3. *Under the GH1 Assumptions, for any $T \in [0,\infty)$, the sequence of processes $\{\lambda_n^{-\frac{1}{2}}\big(A(\lambda_{n,B,2}t) - \lambda_{n,B,2}t\big)_{0\leq t\leq T}, n \geq 1\}$ converges weakly to $\aleph(t)_{0\leq t\leq T}$ 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 [29, Theorem 7.2.3] and formally proven in [142, Theorem 2].

Theorem 4.4. *There exists a w.p.1 continuous Gaussian process* $\mathcal{D}(t)_{t\geq 0}$ *s.t.* $\mathbb{E}[\mathcal{D}(t)] = 0, \mathbb{E}[\mathcal{D}(s)\mathcal{D}(t)] = \mathbb{E}[\big(N_1(s) - s\big)\big(N_1(t) - t\big)]$ *for all* $s, t \geq 0$. *Furthermore, under the GH1 Assumptions, for any* $T \in [0,\infty)$*, the sequence of processes* $\{n^{-\frac{1}{2}}\big(\sum_{i=1}^{n} N_i(t) - nt\big)_{0\leq t\leq T}, n \geq 1\}$ *converges weakly to* $\mathcal{D}(t)_{0\leq t\leq T}$ *in the space* $D[0,T]$ *under the* J_1 *topology.*

Let $\mathcal{Z}_\infty(t)_{t\geq 0}$ denote the Gaussian process s.t. $\mathcal{Z}_\infty(t) = \aleph(t) - \mathcal{D}(t)$ for all $t \geq 0$, and $\mathcal{Z}_{\infty,B}(t)_{t\geq 0}$ denote the Gaussian process s.t. $\mathcal{Z}_{\infty,B}(t) = \aleph(t) - \mathcal{D}(t) - Bt$ for all $t \geq 0$. Existence and continuity of both these processes follows from Theorems 4.3 and 4.4, which further imply the following (as similarly noted in [63]).

Corollary 4.1. *Under the GH1 Assumptions, for any* $T \in [0,\infty)$*, the sequence of processes* $\{n^{-\frac{1}{2}}\big(A(\lambda_{n,B,2}t) - \sum_{i=1}^{n} N_i(t)\big)_{0\leq t\leq T}, n \geq 1\}$ *converges weakly to* $\mathcal{Z}_{\infty,B}(t)_{0\leq t\leq T}$ *in the space* $D[0,T]$ *under the* J_1 *topology.*

4.3.2 Preliminary large deviations results.

Next, 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, and there is a significant literature studying the large deviations of such processes (e.g. [143]). We will rely heavily on the following result, proven in [143] Proposition 1, describing the large deviation behavior of the supremum of certain Gaussian processes. We note that a special case of the same result, customized to the light-tail setting, was also used in [63]. Before stating the result, let us recall the definition of a regularly varying function.

Definition 4.1 (Regularly varying function)**.** *A function $f : \mathcal{R}^+ \to \mathcal{R}^+$ is regularly varying with index γ if for all $t > 0$, $\lim_{x\to\infty} \frac{f(tx)}{f(x)} = t^\gamma$.*

We note that as the complimentary c.d.f.s of heavy-tailed distributions are typically regularly varying, the analysis of regularly varying functions is pervasive in the study of heavy-tailed phenomena, and we refer the interested reader to [144] for an excellent overview of the subject. Then the aforementioned large deviations result is as follows.

Lemma 4.1 ([143] Proposition 1)**.** *Suppose $\mathcal{G}(t)_{t\geq 0}$ is a centered, continuous Gaussian process with stationary increments, satisfying the following conditions.*

- *The associated variance function $\mathbb{E}[\mathcal{G}^2(t)]$ is continuous (on $\mathcal{R}^+$) and regularly varying with index $2H$ for some $0 < H < 1$.*
- *There exists $\epsilon > 0$ s.t. $\lim_{t\downarrow 0} \mathbb{E}[\mathcal{G}^2(t)]|\log(t)|^{1+\epsilon} < \infty$.*

Then for all $\beta > H$ and $c > 0$,

$$\lim_{x\to\infty} \left(\frac{\mathbb{E}[\mathcal{G}^2(x^{\frac{1}{\beta}})]}{x^2} \log \mathbb{P}(\sup_{t\geq 0} \mathcal{Z}(t) - ct^\beta \geq x) \right) = -\frac{1}{2} c^{\frac{2H}{\beta}} (\frac{H}{\beta - H})^{-\frac{2H}{\beta}} (\frac{\beta}{\beta - H})^2.$$

We now use Lemma 4.1 to analyze the large deviations behavior of $\mathcal{Z}_{\infty,B}(t)_{t\geq 0}$, by proving that $\mathcal{Z}_\infty(t)_{t\geq 0}$ satisfies the conditions of Lemma 4.1 for an appropriate parameter of regular variation. The proof relies on certain known results regarding variance of heavy-tailed renewal processes (cf. [131]). In particular, we recall a useful result regarding the variance of heavy-tailed renewal processes. Such results have been proven under considerable generality (e.g. even when the first moment does not exist, and for asymptotic scaling beyond the second moment), although here we state the result customized to our own purposes and assumptions.

Lemma 4.2 ([131])**.** *Under the GH1 assumptions,*

$$\lim_{t\to\infty} \frac{Var[N_1(t)]}{t^{3-\alpha_S}} = 2\big((\alpha_S-1)(2-\alpha_S)(3-\alpha_S)\big)^{-1} C_S.$$

With Lemma 4.2 in hand, we now prove that $\mathcal{Z}_\infty(t)_{t\geq 0}$ satisfies the conditions of Lemma 4.1 for an appropriate parameter of regular variation, deferring all proofs to the appendix.

Lemma 4.3. *Under the GH1 Assumptions, $\mathcal{Z}_\infty(t)_{t\geq 0}$ satisfies the conditions of Lemma 4.1, where $\mathbb{E}[\mathcal{Z}^2(t)]$ is regularly varying with index $3-\alpha_S$.*

Finally, we combine Lemmas 4.1 - 4.3 to prove the desired large deviation results for $\mathcal{Z}_{\infty,B}(t)_{t\geq 0}$, again deferring the proof to the appendix.

Lemma 4.4. *Under the GH1 Assumptions,*

$$\lim_{x\to\infty} \left(x^{1-\alpha_S} \log\left(\mathbb{P}\big(\sup_{t\geq 0} \mathcal{Z}_{\infty,B}(t) \geq x \big) \right) \right) = C_{B,S}; \tag{4.2}$$

Next, we state an additional large deviation-type result, which corresponds to the probability that $\mathcal{Z}_{\infty,B}$ exceeds a large value at the single time at which it is most likely to exceed that value (which will connect to an appropriate lower bound for multi-server queues). The utility of considering such a quantity, in conjunction with the classical notion of large deviations considered in Lemma 4.4, is well-known in the large-deviations literature, and we refer the interested reader to [63] for further discussion. We again defer all proofs to the appendix.

Lemma 4.5. *Under the GH1 Assumptions,*

$$\lim_{x\to\infty} \left(x^{1-\alpha_S} \log\left(\sup_{t\geq 0} \mathbb{P}\big(\mathcal{Z}_{\infty,B}(t) \geq x \big) \right) \right) = C_{B,S}. \tag{4.3}$$

4.3.3 Connecting the large deviations of the limit to the limit of the large deviations, and proof of Theorem 4.1

We now connect the large deviations of $\mathcal{Z}_{\infty,B}$ to the large deviations of $n^{-\frac{1}{2}}\big(A(\lambda_{n,B,2}t) - \sum_{i=1}^{n} N_i(t)\big)$ (for large n, in an appropriate sense) by proving that $\{n^{-\frac{1}{2}} \sup_{t\geq 0} \big(A(\lambda_{n,B,2}t) - \sum_{i=1}^{n} N_i(t)\big), n \geq 1\}$ converges weakly to $\sup_{t\geq 0} \mathcal{Z}_{\infty,B}(t)$, through the lens of Lemma 3.10. We now complete the proof of Theorem 4.1, noting that our proof proceeds similarly to the proof of the analogous large deviations result (which assumed $\mathbb{E}[S^2] < \infty$) in [63].

Proof of Theorem 4.1. We begin by noting that under the GH1 assumptions $\{n^{-\frac{1}{2}}\big(A(\lambda_{n,B,2}t) - \sum_{i=1}^{n} N_i(t)\big), n \geq 1\}$ satisfies the conditions of Lemma 3.10, with limiting stochastic process $\mathcal{Z}_{\infty,B}$. Indeed, condition (2) follows immediately from our proof of Theorem 3.1. Condition (3) follows from Lemma 4.2, since that lemma (along with the definition of $\mathcal{Z}_{\infty,B}$) implies that $\limsup_{t\to\infty} \frac{Var[\mathcal{Z}_{\infty,B}(t)]}{t^{3-\alpha_S}} < \infty$, which (combined with the strictly negative linear drift of $\mathcal{Z}_{\infty,B}$ and a straightforward argument involving the normal distribution which we omit) implies condition (3). Finally, Condition (4) follows from Corollary 4.1, along with the continuity of the supremum map in the J_1 topology, and the fact that convergence in J_1 implies convergence of all co-ordinate projections corresponding to times t such that w.p.1 the limit process has no jump exactly at time t (which will in this case be all $t \geq 0$) [29]. It thus follows from Lemma 3.10 that $\{n^{-\frac{1}{2}} \sup_{t\geq 0} \big(A(\lambda_{n,B,2}t) - \sum_{i=1}^{n} N_i(t)\big), n \geq 1\}$ converges weakly to $\sup_{t\geq 0} \mathcal{Z}_{\infty,B}(t)$. It follows (e.g. from the Portmanteau Theorem) that for all $x \geq 0$,

$$\limsup_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{2}} \sup_{t\geq 0} \big(A(\lambda_{n,B,2}t) - \sum_{i=1}^{n} N_i(t)\big) \geq x\right) \leq \mathbb{P}\big(\sup_{t\geq 0} \mathcal{Z}_{\infty,B}(t) \geq x\big).$$

The first part of Theorem 4.1 (i.e. the upper bound) then follows by combining with Lemma 4.5 and the stochastic comparison result (4.1).

We now prove the second part of Theorem 4.1, i.e. the lower bound, by first reviewing of

a lower bounds from [63] that complements Theorem 3.3. In [63], The authors also prove that the steady-state queue length can be lower-bounded by a different type of supremum, essentially dual to that (3.3) (with the supremum and probability operators interchanged), which we now state here. Let $Z_{n,B,\alpha}$ be a Poisson r.v. with mean $\lambda_{n,B,\alpha}$.

Theorem 4.5 ([63])**.** *Under the same assumptions as Theorem 3.3, supposing in addition that A is exponentially distributed, it holds that for all* $x \geq 0$*,*

$$\mathbb{P}\big(n^{-\frac{1}{\alpha}}\big(Q^n_{A,S,B,\alpha}(\infty)-n\big) \geq x\big) \geq \mathbb{P}(Z_{n,B,\alpha} \geq n) \times \sup_{t\geq 0} \mathbb{P}\left(n^{-\frac{1}{\alpha}}\big(A(\lambda_{n,B,\alpha}t) - \sum_{i=1}^{n} N_i(t)\big) \geq x\right). \tag{4.4}$$

Thus suppose A is exponentially distributed. Let N be a standard normal r.v. Then it follows from Theorem 4.5 that for all $x \geq 0$ and $t \geq 0$, $\liminf_{n\to\infty} \mathbb{P}\big(n^{-\frac{1}{2}} Q^n_{A,S,B,2}(\infty) > x\big)$ is at least

$$\liminf_{n\to\infty} \mathbb{P}\big(Z_{n,B,2} \geq n\big) \times \liminf_{n\to\infty} \mathbb{P}\left(n^{-\frac{1}{2}}\big(A(\lambda_{n,B,2}t) - \sum_{i=1}^{n} N_i(t)\big) > x\right),$$

which by the convergence of the Poisson to the normal, Corollary 4.1, and the Portmanteau Theorem is at least

$$\mathbb{P}\big(N \geq B\big) \times \mathbb{P}\big(\mathcal{Z}_{\infty,B}(t) > x\big).$$

Taking the supremum over all $t \geq 0$, we conclude that

$$\liminf_{n\to\infty} \mathbb{P}\big(n^{-\frac{1}{2}}(Q^n_{A,S,B,2}(\infty) - n) > x\big) \geq \mathbb{P}\big(N \geq B\big) \times \sup_{t\geq 0} \mathbb{P}\big(\mathcal{Z}_{\infty,B}(t) > x\big). \tag{4.5}$$

Combining with Lemma 4.5 and a straightforward limiting argument (the details of which we omit) then completes the proof. □

4.4 Large deviation under HWR-α assumptions, and proof of Theorem 4.2

Now we prove the large deviation result under the Halfin-Whitt-Reed regime, when inter-arrival times have an asymptotically pure Pareto tail for some $\alpha \in (1,2)$ and service times have at least $1+\epsilon$ moment for some $\epsilon \in (0,1]$, i.e. Theorem 4.2.

Recall the bound from Theorem 3.2 and Corollary 3.2,

$$\begin{aligned} & \limsup_{n\to\infty} \mathbb{P}\bigg(n^{-\frac{1}{\alpha}} (Q^n_{A,S,B,\alpha}(\infty) - n) > x \bigg) \\ \leq \;& \mathbb{P}\bigg(\sup_{t\geq 0} \bigg(-(\frac{C_A}{C_\alpha})^{\frac{1}{\alpha}} \hat{S}_\alpha(t,1,0) - Bt \bigg) > x \bigg) \qquad (4.6) \\ = \;& \exp\big(-(\frac{B}{C_A \alpha \Gamma(-\alpha)})^{\frac{1}{\alpha-1}} x \big). \qquad (4.7) \end{aligned}$$

Then the first part of Theorem 4.4 follows trivially from (4.7). For the second part, and it suffices to find a lower bound that have the exact same exponential rate of decay, under the assumption of deterministic service times. For that we first introduce following result, which follows from the results of [1], where the author actually proved the analogous results for waiting times. We include a formal proof translating those results to the setting of steady-state queue in the appendix.

Theorem 4.6 ([1])**.** *Suppose the HWR-α assumptions hold for some $\alpha \in (1,2)$, and in addition S is deterministic (i.e. the queueing system is a $GI/D/n$ queue). Then there is a dense subset $\mathcal{S}$ of $\mathcal{R}^+$ s.t. for all $x \in \mathcal{S}$,*

$$\lim_{n\to\infty} \mathbb{P}\bigg(n^{-\frac{1}{\alpha}} \big(Q^n_{A,S,B,\alpha}(\infty) - n\big)^+ > x \bigg) = \mathbb{P}\bigg(\sup_{k\geq 0} \bigg(-(\frac{C_A}{C_\alpha})^{\frac{1}{\alpha}} S_\alpha(k,1,0) - Bk \bigg) > x \bigg). \qquad (4.8)$$

Intriguingly, we can observe that weak limit in (4.8) is nearly identical to the (4.6), the only difference being that the supremum is taken over positive integer times, instead of all positive real times. In light of Theorem 4.6, our upper bound (holding for general

service time distributions) is in some sense nearly tight even for the very special case of deterministic service times. Indeed, it is well-known that for a process with stationary and independent increments, there are straightforward ways to neatly bound the gap between the all-time supremum and the supremum over integer times (cf. [145, 146]). We now use such analysis to prove that the large deviations behavior of our upper bound is matched for the special case of deterministic service times, i.e. both exhibit the same exponential rate of decay, completing the proof of Theorem 4.2.

Proof of Theorem 4.2. Our approach is very similar to that used in [145]. Let $X(t) \stackrel{\Delta}{=} -(\frac{C_A}{C_\alpha})^{\frac{1}{\alpha}} S_\alpha(t,1,0) - Bt$. For $x > 0$, let $\tau(x) \stackrel{\Delta}{=} \inf\left\{t \geq 0 : X(t) \geq x\right\}$, with $\tau(x) = \infty$ if the process never reaches a value greater than or equal to x. In that case, for any $x > 0$ and $c \in (0, x)$, it follows from stationary and independent increments, and the strong Markov property, that

$$\begin{aligned}
\mathbb{P}\big(\sup_{t\geq 0} X(t) > x, \sup_{k\geq 0} X(k) \leq x - c\big) &\leq \mathbb{P}\bigg(\tau(x) < \infty, \inf_{s\in[\tau(x),\tau(x)+1]} X(s) - X\big(\tau(x)\big) \leq -c\bigg) \\
&= \mathbb{P}\big(\tau(x) < \infty\big) \times \mathbb{P}\big(\inf_{s\in[0,1]} X(s) \leq -c\big) \\
&= \mathbb{P}\big(\sup_{t\geq 0} X(t) > x\big)\mathbb{P}\big(\inf_{s\in[0,1]} X(s) \leq -c\big). \qquad (4.9)
\end{aligned}$$

Combining with the fact that (by a union bound)

$$\mathbb{P}\big(\sup_{t\geq 0} X(t) > x\big) \leq \mathbb{P}\big(\sup_{k\geq 0} X(k) > x - c\big) + \mathbb{P}\big(\sup_{t\geq 0} X(t) > x, \sup_{k\geq 0} X(k) \leq x - c\big), \qquad (4.10)$$

we conclude that

$$\mathbb{P}\big(\sup_{t\geq 0} X(t) > x\big) \leq \mathbb{P}\big(\sup_{k\geq 0} X(k) > x - c\big) + \mathbb{P}\big(\sup_{t\geq 0} X(t) > x\big)\mathbb{P}\big(\inf_{s\in[0,1]} X(s) \leq -c\big),$$

and thus

$$\mathbb{P}\big(\sup_{t\geq 0} X(t) > x\big) \leq \mathbb{P}\big(\sup_{k\geq 0} X(k) > x - c\big) \times \left(\mathbb{P}\big(\inf_{s\in[0,1]} X(s) > -c\big)\right)^{-1}. \qquad (4.11)$$

As $\inf_{s\in[0,1]} X(s)$ is a.s. finite, we may select c sufficiently large to ensure that $\mathbb{P}\big(\inf_{s\in[0,1]} X(s) > -c\big) > 0$, in which case taking the appropriate limit as $x \to \infty$ completes the proof. $\square$

4.5 Conclusion

In this chapter, we provided the first large deviation analysis of steady-state multi-server queues in the Halfin-Whitt regime when service times have infinite variance. When service times have an asymptotically Pareto tail with index $\alpha \in (1, 2)$, we are able to bound the large deviations behavior of the limiting process (defined as any suitable subsequential limit) and derived a matching lower bound when inter-arrival times are Markovian. Interestingly, we find that the large deviations behavior of the limit has a sub-exponential decay, differing fundamentally from the exponentially decaying tails known to hold in the light-tailed setting. Also, for the setting where instead the inter-arrival times have an asymptotically Pareto tail with index $\alpha \in (1, 2)$, we prove a universal bound on the large deviations behavior of the associated limiting process, and prove that even the setting of deterministic service times yields a matching large deviations exponent.

Our work also leaves several interesting directions for future research. It would be very interesting (to again use stochastic comparison approach) to analyze the large deviations behavior of multi-server queues with heavy-tailed service times for a fixed number of servers, where it is known that the interaction between the number of servers, the traffic intensity, and the large deviations behavior can be very subtle [118]. Also, it will be very interesting to further investigate to what extent the bound on the large deviation exponent introduced in Theorem 4.2 is tight (tightness was only certified for the case of deterministic service times). As we can see, during the proof where tightness and weak limit of

$\left\{n^{-\frac{1}{\alpha}}(Q^n_{A,S,B,\alpha}(\infty)-n), n \geq 1\right\}$ were established, the service time distribution played virtually no roles but were washed away. It would be interesting to consider other special classes of service time distributions, gaining more insights on the actual quality of the bound.

4.6 Appendix

4.6.1 Proof of Lemma 4.3

Proof of Lemma 4.3. That $\mathcal{Z}(t)_{t\geq 0}$ is (w.p.1) continuous, centered, and has the stationary increments property follows from the corresponding properties of $\aleph(t)_{t\geq 0}$ and $\mathcal{D}(t)_{t\geq 0}$. Since

$$\mathbb{E}[\mathcal{Z}^2(t)] = c_A^2 t + Var[N_1(t)], \tag{4.12}$$

continuity of $\mathbb{E}[\mathcal{Z}^2(t)]$, as well as the fact that $\lim_{t\downarrow 0}\mathbb{E}[\mathcal{Z}^2(t)]\log^2(t) = 0$, follows from the integral representation Lemma 3.2. Combining with the regular variation implied by Lemma 4.2 completes the proof. □

4.6.2 Proof of Lemma 4.4

Proof of Lemma 4.4. It follows from Lemma 4.3 that under the GH1 Assumptions, we may apply Lemma 4.1 to $\sup_{t\geq 0}\mathcal{Z}_{\infty,B}(t)$, with $\mathcal{G}(t)_{t\geq 0} = \mathcal{Z}_\infty(t)_{t\geq 0}, c = B, \beta = 1, H = \frac{1}{2}(3-\alpha_S)$. It follows from Lemma 4.2 and (4.12) that (in the language of Lemma 4.1)

$$\lim_{x\to\infty}\left(\left(\frac{\mathbb{E}[\mathcal{G}^2(x^{\frac{1}{\beta}})]}{x^2}\right)x^{\alpha_S-1}\right) = 2\big((\alpha_S-1)(2-\alpha_S)(3-\alpha_S)\big)^{-1}C_S, \tag{4.13}$$

and

$$-\frac{1}{2}c^{\frac{2H}{\beta}}\left(\frac{H}{\beta-H}\right)^{-\frac{2H}{\beta}}\left(\frac{\beta}{\beta-H}\right)^2 = -2B^{3-\alpha_S}(3-\alpha_S)^{-(3-\alpha_S)}(\alpha_S-1)^{-(\alpha_S-1)}. \tag{4.14}$$

Combining with Lemma 4.1 and some straightforward algebra completes the proof. □

4.6.3 Proof of Lemma 4.5

Proof of Lemma 4.5. For $x \in \mathcal{R}^+$, let $T_{S,x} \stackrel{\Delta}{=} \frac{(3-\alpha_S)x}{B(\alpha_S-1)}$, and let G denote a standard normal r.v. Note that

$$\begin{aligned}
&\liminf_{x\to\infty}\left(x^{1-\alpha_S}\log\left(\sup_{t\geq 0}\mathbb{P}\big(\mathcal{Z}_{\infty,B}(t)\geq x\big)\right)\right)\\
&\geq \liminf_{x\to\infty}\left(x^{1-\alpha_S}\log\left(\mathbb{P}\big(\mathcal{Z}_{\infty,B}(T_{S,x})\geq x\big)\right)\right)\\
&= \liminf_{x\to\infty} x^{1-\alpha_S}\log\left(\mathbb{P}\left(G>2(\alpha_S-1)^{-1}x\big(\mathbb{E}[\mathcal{Z}^2_{\infty,B}(T_{S,x})]\big)^{-\frac{1}{2}}\right)\right). \quad (4.15)
\end{aligned}$$

As it follows from Lemma 4.3 that $\lim_{x\to\infty} x\big(\mathbb{E}[\mathcal{Z}^2_{\infty,B}(T_{S,x})]\big)^{-\frac{1}{2}} = \infty$, and standard bounds for the normal distribution c.d.f. (cf. [83] Lemma 6) imply that there exists y_0 s.t. $y > y_0$ implies $\mathbb{P}(G > y) \geq \exp(-\frac{y^2}{2} - y)$, we may further conclude that (4.15) is at least

$$-\liminf_{x\to\infty} x^{1-\alpha_S}\left(2(\alpha_S-1)^{-2}x^2\big(\mathbb{E}[\mathcal{Z}^2_{\infty,B}(T_{S,x})]\big)^{-1}+2(\alpha_S-1)^{-1}x\big(\mathbb{E}[\mathcal{Z}^2_{\infty,B}(T_{S,x})]\big)^{-\frac{1}{2}}\right),$$

which by Lemma 4.2, (4.12), and some straightforward algebra equals $C_{B,S}$. Combining with the fact that, by the basic properties of the supremum operator,

$$\limsup_{x\to\infty}\left(x^{1-\alpha_S}\log\left(\sup_{t\geq 0}\mathbb{P}\big(\mathcal{Z}_{\infty,B}(t)\geq x\big)\right)\right)$$

is bounded (from above) by the left-hand-side of (4.2) completes the proof. □

4.6.4 Proof of Theorem 4.6

In [1], Reed proves the following result.

Theorem 4.7. *Suppose that the HWR-α assumptions hold, and in addition S is deterministic (i.e. the system is $GI/D/n$). Then $\{n^{1-\frac{1}{\alpha}}W^n_{A,S,B,\alpha}(\infty), n > B^{\frac{\alpha}{\alpha-1}}\}$ converges in*

distribution to $\sup_{k\geq 0}\left(-(\frac{C_A}{C_\alpha})^{\frac{1}{\alpha}}S_\alpha(k,1,0)-Bk\right)$.

With Theorem 4.7 in hand, we now apply the distributional Little's Law (and more generally the methodology of [115], which had been applied to the light-tailed setting) to derive the corresponding result for queue-lengths, Theorem 4.6.

Proof of Theorem 4.6. Since the system is FCFS with i.i.d. inter-arrival and service times, and service times are deterministic (and hence there is no over-taking), the Distributional Little's Law applies ([147]), and we have

$$Q^n_{A,S,B,\alpha}(\infty)-n \sim A\Big(\lambda_{n,B,\alpha}\big(1+W^n_{A,S,B,\alpha}(\infty)\big)\Big), \tag{4.16}$$

with $A(t)_{t\geq 0}$ and $W^n_{A,S,B,\alpha}(\infty)$ independent. Let $\{A'_i, i\geq 1\}$ denote the sequence of inter-event times in $A(t)_{t\geq 0}$, i.e. A'_1 is drawn from the equilibrium distribution, and $\{A'_i, i\geq 2\}$ are i.i.d. distributed as A. Then for all $x>0$, $\mathbb{P}\Big(n^{-\frac{1}{\alpha}}\big(Q^n_{A,S,B,\alpha}(\infty)-n\big)>x\Big)$ equals

$$\begin{aligned}
&\mathbb{P}\Big(A\Big(\lambda_{n,B,\alpha}\big(1+W^n_{A,S,B,\alpha}(\infty)\big)\Big)>n+xn^{\frac{1}{\alpha}}\Big)\\
=\;&\mathbb{P}\Big(\sum_{i=1}^{\lceil n+xn^{\frac{1}{\alpha}}\rceil}A'_i\leq \lambda_{n,B,\alpha}\big(1+W^n_{A,S,B,\alpha}(\infty)\big)\Big)\\
=\;&\mathbb{P}\left(\frac{\sum_{i=1}^{\lceil n+xn^{\frac{1}{\alpha}}\rceil}(A'_i-1)}{\lceil n+xn^{\frac{1}{\alpha}}\rceil^{\frac{1}{\alpha}}}\leq \frac{\lambda_{n,B,\alpha}\big(1+W^n_{A,S,B,\alpha}(\infty)\big)-\lceil n+xn^{\frac{1}{\alpha}}\rceil}{\lceil n+xn^{\frac{1}{\alpha}}\rceil^{\frac{1}{\alpha}}}\right)
\end{aligned} \tag{4.17}$$

It follows from Theorem 3.4 that

$$\left\{\frac{\sum_{i=1}^{\lceil n+xn^{\frac{1}{\alpha}}\rceil}(A'_i-1)}{\lceil n+xn^{\frac{1}{\alpha}}\rceil^{\frac{1}{\alpha}}}, n\geq 1\right\} \text{ converges in distribution to } (\frac{C_A}{C_\alpha})^{\frac{1}{\alpha}}S_\alpha(1,1,0). \tag{4.18}$$

Theorem 4.7 implies that

$$\left\{\frac{\lambda_{n,B,\alpha}W^n_{A,S,B,\alpha}(\infty)}{\lceil n+xn^{\frac{1}{\alpha}}\rceil^{\frac{1}{\alpha}}}, n\geq 1\right\} \text{ converges in distribution to } \sup_{k\geq 0}\left(-(\frac{C_A}{C_\alpha})^{\frac{1}{\alpha}}S_\alpha(k,1,0)-Bk\right). \tag{4.19}$$

Also, it is easily verified that

$$\lim_{n\to\infty}\frac{\lambda_{n,B,\alpha}-\lceil n+xn^{\frac{1}{\alpha}}\rceil}{\lceil n+xn^{\frac{1}{\alpha}}\rceil^{\frac{1}{\alpha}}}=-B-x. \tag{4.20}$$

As in [115], it then follows from the independence of $\{A'_i, i\geq 1\}$ and $W^n_{A,S,B,\alpha}(\infty)$, and the CLT for triangular arrays (cf. [148]) that for all x which are continuity points of the c.d.f. of $\sup_{k\geq 1}\left(-(\frac{C_A}{C_\alpha})^{\frac{1}{\alpha}}S_\alpha(k,1,0)-Bk\right)$, it holds that

$$\lim_{n\to\infty}\mathbb{P}\left(n^{-\frac{1}{\alpha}}\left(Q^n_{A,S,B,\alpha}(\infty)-n\right)>x\right)=\mathbb{P}\left(\sup_{k\geq 1}\left(-(\frac{C_A}{C_\alpha})^{\frac{1}{\alpha}}S_\alpha(k,1,0)-Bk\right)>x\right). \tag{4.21}$$

The desired result then follows by applying the max-plus operator to both sides. □

REFERENCES

[1] C. Hurvich, J. Reed, *et al.*, "Series expansions for the all-time maximum of α-stable random walks," *Advances in Applied Probability*, vol. 48, no. 3, pp. 744–767, 2016.

[2] T. Engset, "On the calculation of switches in an automatic telephone system-an investigation regarding some points in the basis for the application of probability theory on the determination of the amount of automatic exchange equipment," *Telektronikk*, vol. 94, pp. 99–142, 1998.

[3] T Engset, "Die wahrscheinlichkeitsrechnung zur bestimmung der wahleranzahl in automatischen fernsprechamtern," *Elektrotechnische zeitschrift*, vol. 39, no. 31, pp. 304–306, 1918.

[4] A. K. Erlang, "The theory of probabilities and telephone conversations," *Nyt Tidsskrift for Matematik B*, vol. 20, no. 33-39, p. 16, 1909.

[5] ——, "Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges," *Elektrotkeknikeren*, vol. 13, pp. 5–13, 1917.

[6] L. Kleinrock, "Queueing systems, volume I: theory," 1975.

[7] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, "Statistical analysis of a telephone call center: a queueing-science perspective," *Journal of the American statistical association*, vol. 100, no. 469, pp. 36–50, 2005.

[8] S. Zeltyn and A. Mandelbaum, "Call centers with impatient customers: many-server asymptotics of the M/M/n+ G queue," *Queueing Systems*, vol. 51, no. 3, pp. 361–402, 2005.

[9] Z. Aksin, M. Armony, and V. Mehrotra, "The modern call center: a multi-disciplinary perspective on operations management research," *Production and operations management*, vol. 16, no. 6, pp. 665–688, 2007.

[10] L. M. Wein, A. H. Wilkins, M. Baveja, and S. E. Flynn, "Preventing the importation of illicit nuclear materials in shipping containers," *Risk Analysis*, vol. 26, no. 5, pp. 1377–1393, 2006.

[11] L. M. Wein, Y. Liu, and A. Motskin, "Analyzing the Homeland Security of the US-Mexico Border," *Risk Analysis*, vol. 29, no. 5, pp. 699–713, 2009.

[12] H. Khazaei, J. Misic, and V. B. Misic, "Performance analysis of cloud computing centers using M/G/m/m+r queuing systems," *IEEE Transactions on parallel and distributed systems*, vol. 23, no. 5, pp. 936–943, 2012.

[13] Y. Hu, J. Wong, G. Iszlai, and M. Litoiu, "Resource provisioning for cloud computing," in *Proceedings of the 2009 Conference of the Center for Advanced Studies on Collaborative Research*, IBM Corp., 2009, pp. 101–111.

[14] H. Khazaei, J. Misic, and V. B. Misic, "Modelling of cloud computing centers using M/G/m queues," in *Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on*, IEEE, 2011, pp. 87–92.

[15] R. Cont and A. De Larrard, "Price dynamics in a markovian limit order market," *SIAM Journal on Financial Mathematics*, vol. 4, no. 1, pp. 1–25, 2013.

[16] R. Cont, S. Stoikov, and R. Talreja, "A stochastic model for order book dynamics," *Operations research*, vol. 58, no. 3, pp. 549–563, 2010.

[17] M. K. Govil and M. C. Fu, "Queueing theory in manufacturing: a survey," *Journal of manufacturing systems*, vol. 18, no. 3, p. 214, 1999.

[18] G. R. Bitran and S. Dasu, "A review of open queueing network models of manufacturing systems," *Queueing systems*, vol. 12, no. 1, pp. 95–133, 1992.

[19] C Lakshmi and S. A. Iyer, "Application of queueing theory in health care: a literature review," *Operations research for health care*, vol. 2, no. 1, pp. 25–39, 2013.

[20] J Preater, "Queues in health," *Health Care Management Science*, vol. 5, no. 4, pp. 283–283, 2002.

[21] H. C. Tijms, M. H. Van Hoorn, and A. Federgruen, "Approximations for the steady-state probabilities in the M/G/c queue," *Advances in Applied Probability*, vol. 13, no. 01, pp. 186–206, 1981.

[22] G. Newell, "The $M/G/\infty$ queue," *SIAM Journal on Applied Mathematics*, vol. 14, no. 1, pp. 86–88, 1966.

[23] J. Cohen and O. Boxma, *Boundary value problems in queueing system analysis. 1983*.

[24] J. Cohen, "On the M/G/2 queueing model," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 231–248, 1982.

[25] U. N. Bhat, M. Shalaby, and M. J. Fischer, "Approximation techniques in the solution of queueing problems," *Naval Research Logistics Quarterly*, vol. 26, no. 2, pp. 311–326, 1979.

[26] W. G. Marchal, "Technical notesome simpler bounds on the mean queuing time," *Operations Research*, vol. 26, no. 6, pp. 1083–1088, 1978.

[27] M. F. Neuts, "Computational uses of the method of phases in the theory of queues," *Computers & Mathematics with Applications*, vol. 1, no. 2, pp. 151–166, 1975.

[28] A. Wall and D. Worthington, "Using discrete distributions to approximate general service time distributions in queueing models," *Journal of the Operational Research Society*, vol. 45, no. 12, pp. 1398–1404, 1994.

[29] W. Whitt, *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer Science & Business Media, 2002.

[30] J. Kingman, "On queues in heavy traffic," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 383–392, 1962.

[31] S. Halfin and W. Whitt, "Heavy-traffic limits for queues with many exponential servers," *Operations research*, vol. 29, no. 3, pp. 567–588, 1981.

[32] D. P. Kennedy, "Rates of convergence for queues in heavy traffic. II: sequences of queueing systems," *Advances in Applied Probability*, vol. 4, no. 02, pp. 382–391, 1972.

[33] S. V. Nagaev, "On the speed of convergence in a boundary problem. I," *Theory of Probability & Its Applications*, vol. 15, no. 2, pp. 163–186, 1970.

[34] ——, "On the speed of convergence in a boundary problem. II," *Theory of Probability & Its Applications*, vol. 15, no. 3, pp. 403–429, 1970.

[35] J. Köllerström, "Heavy traffic theory for queues with several servers. II," *Journal of Applied Probability*, vol. 16, no. 02, pp. 393–401, 1979.

[36] J. Dai, A. Dieker, and X. Gao, "Validity of heavy-traffic steady-state approximations in many-server queues with abandonment," *Queueing Systems*, vol. 78, no. 1, pp. 1–29, 2014.

[37] A. Braverman, J. Dai, *et al.*, "Steins method for steady-state diffusion approximations of $M/Ph/n+M$ systems," *The Annals of Applied Probability*, vol. 27, no. 1, pp. 550–581, 2017.

[38] A. Braverman and J. Dai, "High order steady-state diffusion approximation of the Erlang-C system," *arXiv preprint arXiv:1602.02866*, 2016.

[39] A. Braverman, J. Dai, J. Feng, *et al.*, "Steins method for steady-state diffusion approximations: an introduction through the Erlang-A and Erlang-C models," *Stochastic Systems*, vol. 6, no. 2, pp. 301–366, 2016.

[40] I. Gurvich, J. Huang, and A. Mandelbaum, "Excursion-based universal approximations for the Erlang-A queue in steady-state," *Mathematics of Operations Research*, vol. 39, no. 2, pp. 325–373, 2013.

[41] I. Gurvich *et al.*, "Diffusion models and steady-state approximations for exponentially ergodic Markovian queues," *The Annals of Applied Probability*, vol. 24, no. 6, pp. 2527–2559, 2014.

[42] A. Mandelbaum, W. A. Massey, and M. I. Reiman, "Strong approximations for Markovian service networks," *Queueing Systems*, vol. 30, no. 1, pp. 149–201, 1998.

[43] A. J. E. M. Janssen, J. Van Leeuwaarden, and B. Zwart, "Corrected asymptotics for a multi-server queue in the Halfin-Whitt regime," *Queueing Systems*, vol. 58, no. 4, p. 261, 2008.

[44] A. Janssen, J. S. Van Leeuwaarden, and B. Zwart, "Refining square-root safety staffing by expanding Erlang C," *Operations Research*, vol. 59, no. 6, pp. 1512–1522, 2011.

[45] D. Gamarnik and A. Zeevi, "Validity of heavy traffic steady-state approximations in generalized Jackson networks," *The Annals of Applied Probability*, pp. 56–90, 2006.

[46] D. Gamarnik and P. Momčilović, "Steady-state analysis of a multiserver queue in the Halfin-Whitt regime," *Advances in Applied Probability*, vol. 40, no. 2, pp. 548–577, 2008.

[47] D. Gamarnik and A. L. Stolyar, "Multiclass multiserver queueing system in the Halfin–Whitt heavy traffic regime: asymptotics of the stationary distribution," *Queueing Systems*, vol. 71, no. 1-2, pp. 25–51, 2012.

[48] J. Huang and I. Gurvich, "Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue," 2016.

[49] J. Kingman, "Inequalities in the theory of queues," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 102–110, 1970.

[50] D. J. Daley, "Some results for the mean waiting-time and workload in GI/GI/k queues," *Frontiers in queueing: models and applications in science and engineering*, pp. 35–59, 1997.

[51] R. Loulou, "Multi-channel queues in heavy traffic," *Journal of Applied Probability*, vol. 10, no. 04, pp. 769–777, 1973.

[52] S. Y. Oliver, "Stochastic bounds for heterogeneous-server queues with Erlang service times," *Journal of Applied Probability*, vol. 11, no. 04, pp. 785–796, 1974.

[53] E. Arjas and T. Lehtonen, "Approximating many server queues by means of single server queues," *Mathematics of Operations Research*, vol. 3, no. 3, pp. 205–223, 1978.

[54] S. L. Brumelle, "Some inequalities for parallel-server queues," *Operations Research*, vol. 19, no. 2, pp. 402–413, 1971.

[55] A. Scheller-Wolf, "Necessary and sufficient conditions for delay moments in FIFO multiserver queues with an application comparing s slow servers with one fast one," *Operations Research*, vol. 51, no. 5, pp. 748–758, 2003.

[56] A. Scheller-Wolf and R. Vesilo, "Structural interpretation and derivation of necessary and sufficient conditions for delay moments in FIFO multiserver queues," *Queueing Systems*, vol. 54, no. 3, pp. 221–232, 2006.

[57] R. Vesilo and A. Scheller-Wolf, "Delay moment bounds for multiserver queues with infinite variance service times," *INFOR: Information Systems and Operational Research*, vol. 51, no. 4, pp. 161–174, 2013.

[58] M. Grossglauser and J.-C. Bolot, "On the relevance of long-range dependence in network traffic," *IEEE/ACM Transactions on Networking (TON)*, vol. 7, no. 5, pp. 629–640, 1999.

[59] A. Marazzi, F. Paccaud, C. Ruffieux, and C. Beguin, "Fitting the distributions of length of stay by parametric models," *Medical care*, vol. 36, no. 6, pp. 915–927, 1998.

[60] S. I. Resnick, *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

[61] J. Reed *et al.*, "The G/GI/N queue in the Halfin–Whitt regime," *The Annals of Applied Probability*, vol. 19, no. 6, pp. 2211–2269, 2009.

[62] A. A. Puhalskii, J. E. Reed, *et al.*, "On many-server queues in heavy traffic," *The Annals of Applied probability*, vol. 20, no. 1, pp. 129–195, 2010.

[63] D. Gamarnik, D. A. Goldberg, *et al.*, "Steady-state GI/G/n queue in the HalfinWhitt regime," *The Annals of Applied Probability*, vol. 23, no. 6, pp. 2382–2419, 2013.

[64] A. J. E. M. Janssen and J. Van Leeuwaarden, "Back to the roots of the M/D/s queue and the works of Erlang, Crommelin and Pollaczek," *Statistica Neerlandica*, vol. 62, no. 3, pp. 299–313, 2008.

[65] D Worthington, "Reflections on queue modelling from the last 50 years," *Journal of the Operational Research Society*, vol. 60, no. 1, S83–S92, 2009.

[66] D. G. Kendall, "Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain," *The Annals of Mathematical Statistics*, pp. 338–354, 1953.

[67] ——, "Some problems in the theory of queues," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 151–185, 1951.

[68] R. Aghajani and K. Ramanan, "The limit of stationary distributions of many-server queues in the halfin-whitt regime," *arXiv preprint arXiv:1610.01118*, 2016.

[69] J. Kingman, "The heavy traffic approximation in the theory of queues," in *Proceedings of the Symposium on Congestion Theory*, University of North Carolina Press, Chapel Hill, NC, 1965.

[70] J. Köllerström, "Heavy traffic theory for queues with several servers. I," *Journal of Applied Probability*, vol. 11, no. 03, pp. 544–552, 1974.

[71] A. Borovkov, "Some limit theorems in the theory of mass service, II multiple channels systems," *Theory of Probability & Its Applications*, vol. 10, no. 3, pp. 375–400, 1965.

[72] D. L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic. II: sequences, networks, and batches," *Advances in Applied Probability*, vol. 2, no. 02, pp. 355–369, 1970.

[73] V. Gupta, M. Harchol-Balter, J. Dai, and B. Zwart, "On the inapproximability of M/G/K: why two moments of job size distribution are not enough," *Queueing Systems*, vol. 64, no. 1, pp. 5–48, 2010.

[74] V. Gupta and T. Osogami, "On markov–krein characterization of the mean waiting time in M/G/K and other queueing systems," *Queueing Systems*, vol. 68, no. 3, pp. 339–352, 2011.

[75] T. Rolski and D. Stoyan, "Technical NoteOn the Comparison of Waiting Times in GI/G/1 Queues," *Operations Research*, vol. 24, no. 1, pp. 197–200, 1976.

[76] W. Whitt, "Comparison conjectures about the M/G/s queue," *Operations Research Letters*, vol. 2, no. 5, pp. 203–209, 1983.

[77] ——, "The effect of variability in the GI/G/s queue," *Journal of Applied Probability*, vol. 17, no. 04, pp. 1062–1071, 1980.

[78] Y. Zheng, N. Shroff, and P. Sinha, "Heavy Traffic Limits for GI/H/n Queues: Theory and Application," *arXiv preprint arXiv:1409.3463*, 2014.

[79] S. Chawla, N. R. Devanur, A. E. Holroyd, A. Karlin, J. Martin, and B. Sivan, "Stability of service under time-of-use pricing," *arXiv preprint arXiv:1704.02364*, 2017.

[80] D. R. Smith and W. Whitt, "Resource sharing for efficiency in traffic systems," *Bell System Technical Journal*, vol. 60, no. 1, pp. 39–55, 1981.

[81] R. W. Wolff, "Upper bounds on work in system for multichannel queues," *Journal of applied probability*, vol. 24, no. 02, pp. 547–551, 1987.

[82] M. J. Sobel, "Not–Simple Inequalities for Multiserver Queues," *Management Science*, vol. 26, no. 9, pp. 951–956, 1980.

[83] D. A. Goldberg, "On the steady-state probability of delay and large negative deviations for the GI/GI/n queue in the halfin-whitt regime," *arXiv preprint arXiv:1307.0241*, 2016.

[84] R. Atar and N. Solomon, "Asymptotically optimal interruptible service policies for scheduling jobs in a diffusion regime with nondegenerate slowdown," *Queueing Systems*, vol. 69, no. 3, pp. 217–235, 2011.

[85] J. Blanchet, J. Dong, and Y. Pei, "Perfect sampling of GI/GI/c queues," *arXiv preprint arXiv:1508.02262*, 2015.

[86] E. C. Ni and S. G. Henderson, "How hard are steady-state queueing simulations?" *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 25, no. 4, p. 27, 2015.

[87] D. Bertsimas, "An analytic approach to a general class of G/G/s queueing systems," *Operations Research*, vol. 38, no. 1, pp. 139–155, 1990.

[88] V Ramaswami and D. M. Lucantoni, "Algorithms for the multi-server queue with phase type service," *Stochastic Models*, vol. 1, no. 3, pp. 393–417, 1985.

[89] J. Dai, S. He, *et al.*, "Many-server queues with customer abandonment: numerical analysis of their diffusion model," *Stochastic Systems*, vol. 3, no. 1, pp. 96–146, 2013.

[90] D. Bertsimas and K. Natarajan, "A semidefinite optimization approach to the steady-state analysis of queueing systems," *Queueing Systems*, vol. 56, no. 1, pp. 27–39, 2007.

[91] C. Bandi, D. Bertsimas, and N. Youssef, "Robust queueing theory," *Operations Research*, vol. 63, no. 3, pp. 676–700, 2015.

[92] A. Doig, "A bibliography on the theory of queues," *Biometrika*, pp. 490–514, 1957.

[93] G. C. Ovuworie, "Multi-channel queues: a survey and bibliography," *International Statistical Review/Revue Internationale de Statistique*, pp. 49–71, 1980.

[94] W. Whitt, "Approximations for the GI/G/m queue," *Production and Operations Management*, vol. 2, no. 2, pp. 114–161, 1993.

[95] J. G. Shanthikumar, "Bounds and an approximation for single server queues," *Journal of the Operations Research Society of Japan*, vol. 26, no. 2, pp. 118–134, 1983.

[96] W. Whitt, "A review of L = λW and extensions," *Queueing Systems*, vol. 9, no. 3, pp. 235–268, 1991.

[97] D. Bertsimas and D. Nakazato, "The distributional Little's law and its applications," *Operations Research*, vol. 43, no. 2, pp. 298–310, 1995.

[98] M. Longnecker and R. Serfling, "General moment and probability inequalities for the maximum partial sum," *Acta Mathematica Hungarica*, vol. 30, no. 1-2, pp. 129–133, 1977.

[99] P. Billingsley, *Convergence of probability measures*. John Wiley & Sons, 2013.

[100] W. Szczotka, "Tightness of the stationary waiting time in heavy traffic," *Advances in Applied Probability*, pp. 788–794, 1999.

[101] W. Szczotka and W. Woyczynski, "Heavy-tailed dependent queues in heavy traffic," *PROBABILITY AND MATHEMATICAL STATISTICS-WROCLAW UNIVERSITY*, vol. 24, no. 1, p. 67, 2004.

[102] A. Gut, *Stopped random walks*. Springer, 2009.

[103] Y. S. Chow, C. A. Hsiung, T. L. Lai, *et al.*, "Extended renewal theory and moment convergence in Anscombe's theorem," *The Annals of Probability*, vol. 7, no. 2, pp. 304–318, 1979.

[104] P. Hitczenko, "Best constants in martingale version of Rosenthal's inequality," *The Annals of Probability*, pp. 1656–1668, 1990.

[105] Y.-F. Ren and H.-Y. Liang, "On the best constant in Marcinkiewicz–Zygmund inequality," *Statistics & probability letters*, vol. 53, no. 3, pp. 227–233, 2001.

[106] T Figiel, P Hitczenko, W Johnson, G Schechtman, and J Zinn, "Extremal properties of Rademacher functions with applications to the Khintchine and Rosenthal inequalities," *Transactions of the American Mathematical Society*, vol. 349, no. 3, pp. 997–1027, 1997.

[107] W. B. Johnson, G. Schechtman, and J. Zinn, "Best constants in moment inequalities for linear combinations of independent and exchangeable random variables," *The Annals of Probability*, pp. 234–253, 1985.

[108] M. Olvera-Cravioto, J. Blanchet, P. Glynn, *et al.*, "On the transition from heavy traffic to heavy tails for the M/G/1 queue: the regularly varying case," *The Annals of Applied Probability*, vol. 21, no. 2, pp. 645–668, 2011.

[109] D. Gamarnik, D. A. Goldberg, *et al.*, "On the rate of convergence to stationarity of the M/M/N queue in the HalfinWhitt regime," *The Annals of Applied Probability*, vol. 23, no. 5, pp. 1879–1912, 2013.

[110] A. Erlang, "On the rational determination of the number of circuits," *The life and works of AK Erlang*, pp. 216–221, 1948.

[111] D. L. Jagerman, "Some properties of the Erlang loss function," *Bell Labs Technical Journal*, vol. 53, no. 3, pp. 525–551, 1974.

[112] R. Aghajani and K. Ramanan, "Ergodicity of an spde associated with a many-server queue," *arXiv preprint arXiv:1512.02929*, 2015.

[113] Y. Li and D. A. Goldberg, "Simple and explicit bounds for multi-server queues with universal $\frac{1}{1-\rho}$ scaling," *preprint*, 2017.

[114] A.-L. Barabasi, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, no. 7039, pp. 207–211, 2005.

[115] P. Jelenković, A. Mandelbaum, and P. Momčilović, "Heavy traffic limits for queues with many deterministic servers," *Queueing Systems*, vol. 47, no. 1, pp. 53–69, 2004.

[116] O. J. Boxma and J. Cohen, "The M/G/1 queue with heavy-tailed service time distribution," *IEEE journal on selected areas in communications*, vol. 16, no. 5, pp. 749–763, 1998.

[117] O. J. Boxma and J. W. Cohen, "Heavy-traffic analysis for the GI/G/1 queue with heavy-tailed distributions," *Queueing systems*, vol. 33, no. 1, pp. 177–204, 1999.

[118] S. Foss and D. Korshunov, "On large delays in multi-server queues with heavy tails," *Mathematics of Operations Research*, vol. 37, no. 2, pp. 201–218, 2012.

[119] S. Foss and D. Korshunov, "Heavy tails in multi-server queue," *Queueing Systems*, vol. 52, no. 1, pp. 31–48, 2006.

[120] J. Blanchet and K. R. Murthy, "Tail asymptotics for delay in a half-loaded gi/gi/2 queue with heavy-tailed job sizes," *Queueing Systems*, vol. 81, no. 4, pp. 301–340, 2015.

[121] W. Whitt, "The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting-time distribution," *Queueing Systems*, vol. 36, no. 1, pp. 71–87, 2000.

[122] G Samorodnitsky and M. Taqqu, "Stable non-gaussian random processes. 1994," *Chapman&Hall, New York*, vol. 29,

[123] N. H. Bingham, "Fluctuation theory in continuous time," *Advances in Applied Probability*, vol. 7, no. 04, pp. 705–766, 1975.

[124] D. J. Daley, "Bounds for the variance of certain stationary point processes," *Stochastic Processes and their Applications*, vol. 7, no. 3, pp. 255–264, 1978.

[125] D. Daley, "Tight bounds for the renewal function of a random walk," *The Annals of Probability*, pp. 615–621, 1980.

[126] G. Lorden, "On excess over the boundary," *The Annals of Mathematical Statistics*, pp. 520–527, 1970.

[127] A Baltrunas and E. Omey, "Second-order renewal theorem in the finite-means case," *Theory of Probability & Its Applications*, vol. 47, no. 1, pp. 127–132, 2003.

[128] J Geluk, "A renewal theorem in the finite-mean case," *Proceedings of the American mathematical society*, vol. 125, no. 11, pp. 3407–3413, 1997.

[129] N. Mohan *et al.*, "Teugels' renewal theorem and stable laws," *The Annals of Probability*, vol. 4, no. 5, pp. 863–868, 1976.

[130] J. L. Teugels, "Renewal theorems when the first or the second moment is infinite," *The Annals of Mathematical Statistics*, vol. 39, no. 4, pp. 1210–1219, 1968.

[131] R. Gaigalas, I. Kaj, *et al.*, "Convergence of scaled renewal processes and a packet arrival model," *Bernoulli*, vol. 9, no. 4, pp. 671–703, 2003.

[132] R. Farrell, "Limit theorems for stopped random walks," *The Annals of Mathematical Statistics*, pp. 1332–1343, 1964.

[133] J. Kingman, "Some inequalities for the queue GI/G/1," *Biometrika*, vol. 49, no. 3/4, pp. 315–324, 1962.

[134] J. A. Bucklew, *Large deviation techniques in decision, simulation, and estimation*. Wiley New York, 1990.

[135] D. W. Stroock, *An introduction to the theory of large deviations*. Springer Science & Business Media, 2012.

[136] A. Shwartz and A. Weiss, *Large deviations for performance analysis: queues, communication and computing*. CRC Press, 1995, vol. 5.

[137] N. G. Duffield and N. O'connell, "Large deviations and overflow probabilities for the general single-server queue, with applications," in *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge Univ Press, vol. 118, 1995, pp. 363–374.

[138] D. Bertsimas, I. C. Paschalidis, J. N. Tsitsiklis, *et al.*, "On the large deviations behavior of acyclic networks of G/G/1 queues," *The Annals of Applied Probability*, vol. 8, no. 4, pp. 1027–1069, 1998.

[139] C.-S. Chang, "Sample path large deviations and intree networks," *Queueing Systems*, vol. 20, no. 1, pp. 7–36, 1995.

[140] A Ganesh and V. Anantharam, "Stationary tail probabilities in exponential server tandems with renewal arrivals," *Queueing Systems*, vol. 22, no. 3, pp. 203–247, 1996.

[141] R. J. Adler, "An introduction to continuity, extrema, and related topics for general gaussian processes," *Lecture Notes-Monograph Series*, vol. 12, pp. i–155, 1990.

[142] W. Whitt, "Queues with superposition arrival processes in heavy traffic," *Stochastic processes and their applications*, vol. 21, no. 1, pp. 81–91, 1985.

[143] A. Dieker, "Conditional limit theorems for queues with gaussian input, a weak convergence approach," *Stochastic processes and their applications*, vol. 115, no. 5, pp. 849–873, 2005.

[144] N. H. Bingham, C. M. Goldie, and J. L. Teugels, *Regular variation*. Cambridge university press, 1989, vol. 27.

[145] K. Maulik and B. Zwart, "Tail asymptotics for exponential functionals of Lévy processes," *Stochastic Processes and their Applications*, vol. 116, no. 2, pp. 156–177, 2006.

[146] E. Willekens, "On the supremum of an infinitely divisible process," *Stochastic processes and their applications*, vol. 26, pp. 173–175, 1987.

[147] R. Haji and G. F. Newell, "A relation between stationary queue and waiting time distributions," *Journal of Applied Probability*, vol. 8, no. 03, pp. 617–620, 1971.

[148] K. L. Chung, *A course in probability theory*. Academic press, 2001.