Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Steady-State Analysis for Multiserver Queues Under Size Interval Task Assignment in the Quality-Driven Regime

Bo Zhanghttp://www2.isye.gatech.edu/~bzhang34/, Bert Zwart,

To cite this article:

Full terms and conditions of use: http://pubsonline.informs.org/page/terms-and-conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article’s accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2013, INFORMS

Please scroll down for article—it is on subsequent pages
Steady-State Analysis for Multiserver Queues Under Size Interval Task Assignment in the Quality-Driven Regime

Bo Zhang
IBM T. J. Watson Research Center, Yorktown Heights, New York, 10598,
bozhang@gatech.edu, http://www2.isye.gatech.edu/~bzhang34/

Bert Zwart
Centrum Wiskunde & Informatica, PNA2, Amsterdam, North-Holland 1098XG, The Netherlands,
bert.zwart@cwi.nl

We study the steady-state behavior of multiserver queues with general job size distributions under size interval task assignment (SITA) policies. Assuming Poisson arrivals and the existence of the $\alpha$th moment of the job size distribution for some $\alpha > 1$, we show that if the job arrival rate and the number of servers increase to infinity with the traffic intensity held fixed, the SITA policy parameterized by $\alpha$ minimizes in a large deviation sense the steady-state probability that the total number of jobs in the system is greater than or equal to the number of servers. The optimal large deviation decay rate can be arbitrarily close to the one for the corresponding probability in an infinite-server queue, which only depends on the system traffic intensity but not on any higher moments of the job size distribution. This supports in a many-server asymptotic framework the common wisdom that separating large jobs from small jobs protects system performance against job size variability.

Key words: multiserver queues; quality-driven regime; size-based task assignment; size interval task assignment policy; large deviations; infinite-server queues

MSC2000 subject classification: Primary: 60F10, 60K25, 68M20; secondary: 60K30, 90B22

OR/MS subject classification: Primary: queues; secondary: limit theorems

History: Received January 27, 2010; revised April 1, 2012. Published online in Articles in Advance November 28, 2012.

1. Introduction. Multiserver queues with Poisson arrivals and general job size distributions are probably the simplest, yet reasonable models for many real-world service systems, such as server farms and call centers. One of the oldest and most fundamental problems arising in this type of systems is the choice of a good rule for assigning jobs (or tasks) to servers, known as the task assignment policy. In general, it is rather difficult to evaluate the performance of almost any task assignment policy in such systems, and thus many basic questions remain open on this subject.

One intriguing open question that motivates our study is regarding the performance of size-based task assignment, which broadly refers to the practice of dispatching jobs of different sizes or different size distributions to separate server pools. The question is whether or not separating large jobs from small ones protects system performance against high variability of job sizes. Intuitively, such a task assignment policy prevents large jobs from occupying all servers and blocking many small jobs. This advantage should become especially pronounced as the job size variability increases. On the other hand, dedicating servers to large and small jobs, respectively, can lead to server idling and underutilization, and thus hurt system performance. This suggests that the right answer to the proposed question should depend on the context, and the specifics of the policy matter.

In this paper, we focus on one type of size-based task assignment policy; namely, the size interval task assignment (SITA) policy. In words, the SITA policy groups all servers into multiple pools, divides the possible job sizes into different intervals, and assigns all jobs in each size interval to one server pool exclusively, with each subsystem operating on a first-come-first-served (FCFS) basis. In §2, we shall provide a more precise definition of SITA.

Our goal is to study, for a system where SITA has been chosen as the scheduling discipline a priori, how to parameterize the SITA policy to minimize the occurrence of congestion, when the number of servers is large and the traffic intensity in the system is bounded away from 1.

Specifically, we study the steady-state performance of the proposed SITA policies in the quality-driven (QD) regime, where the arrival rate and the number of servers $N$ increase to infinity with the traffic intensity ($< 1$) held fixed. Loosely speaking, the QD regime is a many-server, overstaffing regime with a constant overstaffing factor, and it is appropriate for modeling large-scale systems with a nonnegligible amount of slack capacity. The practice of reserving extra capacity is often seen in service-oriented systems, such as emergency call centers, and a variety of computer systems where capacity over provisioning is commonplace to meet high quality of service standards (see López-Ortiz [21] and the references therein).

The performance measure that we are interested in is $\mathbb{P}(Q^N \geq N)$, in particular, its rate of decay in the QD regime, where $Q^N$ denotes the steady-state total number of jobs in the $N$-server system. This probability is a good...
indication of system performance, especially in the QD regime. First, the occurrence of the event \( \text{number of jobs in the system} \geq \text{number of servers} \) implies that there are jobs being delayed. In practice, this could suggest a greater chance of hardware failure or incur extra operational cost due to using spare or outsourced capacity (so-called pay per use; see Liu and Wee [20]). In addition, in the QD regime, \( \mathbb{P}(Q_N \geq N) \) should become a very remote tail probability as \( N \) increases to infinity, which makes it a natural and classical quantity of interest from the point of view of large deviations theory and extreme value theory (see Embrechts et al. [6], Finkenstadt and Rootzen [7]). Furthermore, we believe that studying the asymptotics of \( \mathbb{P}(Q_N \geq N) \) is an important step toward a complete understanding of the distribution of \( Q_N \) and other steady-state performance metrics.

The performance benchmark that we use is the lower bound \( \mathbb{P}(Q_N^* \geq N) \), where \( Q_N^* \) denotes the steady-state number of jobs in the corresponding \( M/G/\infty \) queue. In other words, we are interested in how close to \( \mathbb{P}(Q_N^* \geq N) \) a cleverly chosen SITA policy can make \( \mathbb{P}(Q_N \geq N) \).

More specifically, for any \( \epsilon > 0 \), we shall prescribe the SITA policy under which \( \mathbb{P}(Q_N \geq N) \) is \( \epsilon \)-close to \( \mathbb{P}(Q_N^* \geq N) \) on a logarithmic scale in the QD regime. We hasten to emphasize that SITA may not be the only type of task assignment policy, which can be configured to attain such performance level. In fact, if the job size distribution is exponential or deterministic, exact asymptotic coincidence between \( \mathbb{P}(Q_N \geq N) \) and \( \mathbb{P}(Q_N^* \geq N) \) is even achieved by the simple FCFS policy (see Theorem 5 and Remark 1). However, under a general job size distribution assumption, no other policy has been shown to perform as well or better than our proposed SITA in this setting. In fact, existing results in the literature leave open the possibility that a work-conserving policy like FCFS may not behave well with respect to \( \mathbb{P}(Q_N^* \geq N) \) when the job size variability level is high.

For example, Smith and Whitt [25, §IV], demonstrate that as job size variability increases, the long-run average total number of jobs in a Poisson input two-server queue with a hyperexponential job size can go to infinity under the FCFS policy, while the same system under a size-based task assignment policy that essentially divides the system into two independent \( M/M/1 \) queues maintains a constant long-run average total number of jobs in the system independent of the variability level. The construction of their example is illuminating: if all servers (in their case, both of the two servers) are busy serving large-size jobs occurs, subsequently arriving jobs all have to wait for a long time to be served. If job size variability level is high and thus large jobs are really huge, the clogging effect due to this event of huge jobs occupying all servers can become so pronounced that it dominates the rarity of the event, and eventually hurts long-run average performance badly.

It is worth noting, though, that the above observation is made in a nonasymptotic setting. As the number of servers grows to infinity, i.e., in a many-server asymptotic regime such as QD, one may suspect that the clogging event, on the contrary to the above observation, can become so rare that its impact on long-run system performance becomes negligible. In turn, one may conjecture that a work-conserving discipline such as FCFS should do well in a many-server asymptotic setting, e.g., achieve the same (or even a stronger) form of logarithmic optimality as SITA in the QD regime. Whether the conjecture holds true and, if it does, whether additional assumptions on the job size distribution (or variability level) are needed are interesting, yet challenging questions for further study. Rather than attempt to answer all these questions, we focus on SITA, which, by design, does not allow large jobs to occupy all servers, and obtain the first analytical result on a steady-state performance metric for a scheduling policy in the QD regime.

There are strong motivations to study SITA. First formally introduced by Harchol-Balter et al. [15], the SITA policy has since attracted a lot of research attention (e.g., Bachmat and Sarfati [2], Cardellini et al. [4], Ciardo et al. [5], Harchol-Balter et al. [16, 17], Schroeder and Harchol-Balter [24]) especially in the computer systems performance evaluation community. In Harchol-Balter et al. [15, §3], they state that multiserver queues under SITA were inspired by and used as an abstraction of the \( \text{xolas} \) distributed computing facility at MIT’s Laboratory for Computer Science. Harchol-Balter et al. [15] also argue that the assumption that task sizes are known holds to an approximate degree in other contexts, such as some batch computing servers. Ciardo et al. [5] apply SITA to web server farms, and they point out that if the exact size of each job (i.e., URL request in that setting) is not available to the front-end dispatcher, SITA still can be implemented by a two-stage allocation policy. Schroeder and Harchol-Balter [24] apply SITA to heavy-tailed supercomputing work loads and they argue that in many distributed servers, task assignment is done by the users (rather than a dispatcher); specifically, in the SITA case, each job is submitted with an estimated runtime and different host machines have different duration limitations: up to two hours, up to four hours, up to eight hours, or unlimited, etc. Cardellini et al. [4] also discuss using SITA for web server farms, especially when the Web content is static. For a more complete list of references and a careful discussion on the existing results, we refer the reader to Harchol-Balter et al. [16, §2]. In addition, we note that size-based scheduling policies, such as shortest job first, preemptive shortest job first, and shortest remaining processing time first, are well-studied subjects in the literature of single-server queues (see Harchol-Balter [14], Nuyens et al. [22]).
In terms of the performance of SITA or more generally size-based task assignment, both positive and negative results have been reported. As alluded to above, these results are often provided through a comparison with FCFS. In contrast to Smith and Whitt [25], a recent study by Harchol-Balter et al. [16] shows that the mean job delay in a two-server queue under a size-based policy can diverge, while it converges under FCFS, as the variance of the job size distribution goes to infinity. In fact, Harchol-Balter et al. [16] show that neither size-based task assignment nor FCFS could be a sure win.

To the best of our knowledge, a proof on the superiority (or inferiority) of sized-based task assignment policies has never materialized in any general setting. This lack of analytical results is largely due to the fact that the study of multiserver queues under sized-based task assignment boils down to analyzing several M/G/N subsystems under FCFS in parallel, while the M/G/N queue under FCFS remains an unsolved problem (see Kingman [19]). Most existing results on M/G/N are approximate; for example, see Tijms et al. [29], Wang and Wolff [30], Yao [35], and the references therein. In the many-server asymptotic setting, the only published analytical result on steady-state distributions is provided by Gamarnik and Momčilović [9], who obtain an explicit expression for the critical exponent for the moment generating function of a limiting (scaled) steady-state queue length in the quality-and-efficiency-driven (QED) regime assuming lattice-valued job sizes with a finite support (and thus with a light tail). As for steady-state analysis for many-server queues with heavy-tailed job sizes, we are not aware of any conclusive result. Whitt [31] shows that the steady-state waiting time distribution of the M/G/N queue under FCFS has a heavy tail (with appropriate definition) whenever the job size distribution does. Also, to the best of our knowledge, task assignment has never been studied in a many-server regime.

It is worth emphasizing that extending the analysis to general job size distributions, in particular, with unbounded support, is mathematically challenging. In fact, we are not aware of any steady-state, many-server asymptotic result with a general unbounded job size distribution assumption, except for the very recent work by Gamarnik and Goldberg [8], in which they derive an upper bound on the large deviation exponent of the limiting steady-state FCFS queue length in the QED regime assuming finite \( 2 + \epsilon \)th job size and interarrival time moments. In the aforementioned earlier paper by Gamarnik and Momčilović [9] on the same model, they assume finite support of the job size distribution and develop a novel finite-dimensional Markov chain representation based on this assumption. A recent work by Yang et al. [34] studies SMART scheduling policies in the many sources large deviations regime by developing a two-dimensional queueing framework and the tractability of their framework also relies on the finite support assumption. In this paper, we obtain our main result for the case of a general unbounded job size distribution. In light of the absence of such results, one contribution of our work is to provide some insight into how one can analyze the unbounded support case.

Also, the QD regime has not received much attention in the literature. The service quality in the QD regime is noted to be extremely good for a large number of servers \( N \) (see Garnett et al. [11], Pang and Whitt [23], Zeltyn and Mandelbaum [36]), hence the name. However, it seems that there has almost been no study that attempts to quantify how fast the performance improves as \( N \) increases, which is essential for resource dimensioning. The only work that we are aware of is Zeltyn and Mandelbaum [36], where the performance asymptotics for the M/M/N + G queue in the QD regime are derived based on the exact formulas. Also, we note that estimating the performance of many-server queues by simulation is difficult as well, because it often involves increasingly rare events as \( N \) grows. A recent study in this regard is the paper by Blanchet et al. [3], which develops a rare event simulation algorithm that is asymptotically efficient in the QD regime. It is worth mentioning that there does exist an extensive literature on the analysis of queues where there is a finite number of servers, of which the speed grows, and the number of input processes grows accordingly. This is also known as the many-flows regime, and is motivated by highly multiplexed communication networks (see Ganesh et al. [10] for background). Finally, we note that our benchmark system, the M/G/∞ queue, is a remarkably tractable model and has been exploited to analyze the M/G/N queue in other regimes (e.g., see Whitt [31]).

In short, the main contribution of this paper is the construction of a family of SITA policies under which \( \mathbb{P}(Q^0 \geq N) \) can be made arbitrarily close to \( \mathbb{P}(Q^\infty \geq N) \) in a large deviation sense. Our result holds true for any job size distribution with finite \( \alpha \)th moment for some \( \alpha > 1 \), including those with an infinite variance or a heavy-tailed distribution. To the best of our knowledge, there is no other steady-state, many-server asymptotic result obtained under such a general job size distribution assumption.\(^1\) Furthermore, if the job size has a finite support or a hyperexponential distribution, our proposed size-based task assignment policy achieves an even stronger performance: \( \mathbb{P}(Q^\infty \geq N) \sim \mathbb{P}(Q^\infty \geq N) \) on a logarithmic scale. Because \( \mathbb{P}(Q^\infty \geq N) \) is insensitive to the job size distribution, our results suggest that the recommended size-based task assignment policies indeed

\(^1\) The job arrival processes in Gamarnik and Goldberg [8] and Gamarnik and Momčilović [9] are much more general than the Poisson process in our model.
protect system performance against high job size variability in the QD regime. In the general job size distribution case, the number of size intervals (or equivalently, the number of server pools) in our SITA prescription grows to infinity at sublinear rate as \( N \) increases, and both the size cutoff values and the number of servers allocated to each pool are carefully parameterized by the job size moment index \( \alpha \). These features enable our policies to perform well even when the job size has an unbounded support. Our approach to analyzing systems with many servers and job sizes of unbounded support is potentially interesting for other problems, such as the behavior of the same probability under FCFS in the QD regime, which is open. In obtaining the main results, we also develop some estimation results on random walks and the M/D/N queue, both asymptotics and bounds, which may be of independent interest. In particular, there is a connection with light traffic analysis of random walks (see Asmussen [1]).

In the next section, we formulate our model, provide the organization of the remainder of this paper, and then state our main results in more detail.

2. Model formulation and main results. In this section, we first provide the model formulation, including the definitions of the QD limiting regime and the SITA policy. Then, we summarize the main results of this paper.

We study the M/G/N queue in the QD regime, which is achieved by considering a sequence of queues indexed by the number of servers \( N \) where the arrival rate to the system grows large proportionally to \( N \) and the traffic intensity remains fixed. Specifically, first let \( \{A(t), t \geq 0\} \) be a Poisson process with rate \( \lambda \), and \( T_i, i \geq 1 \) be the interarrival times. In the \( N \)th system, let the job arrival process; namely, \( \{A^N(t), t \geq 0\} \), be such that \( A^N(t) = A(Nt) \) for all \( t \geq 0 \), or equivalently, the interarrival times \( T^N_i, i \geq 1 \) are equal to \( T_i/N \). Denoting by \( \lambda^N \) the arrival rate to the \( N \)th system, we then have \( \lambda^N = \lambda N \). Both the job size distribution and the traffic intensity are the same for any \( N \). Specifically, job sizes are i.i.d. equal in distribution to \( S \), with \( \mathbb{E}[S] = \mu^{-1} \), and each server processes jobs at unit rate. As a result, the traffic intensity in the \( N \)th system is just \( \lambda^N/N\mu = \lambda/\mu \), and we denote this fixed traffic intensity by \( \rho := \lambda/\mu \). The QD regime is achieved by letting \( N \to \infty \). We further assume \( \mathbb{P}[S > 0] = 1 \) for simplicity. This assumption is not restrictive, because if \( \mathbb{P}[S > 0] < 1 \), one may just allocate one server for zero-size jobs and apply our proposed policies, with \( N \) replaced by \( N - 1 \), to the other servers and jobs, leading to the same optimality results established in this paper.

As a performance benchmark for the sequence of multiserver queues in the QD regime, we also consider a sequence of infinite-server queues: in the \( N \)th system, jobs with i.i.d. sizes equal in distribution to \( S \) arrive according to the process \( \{A^N(t), t \geq 0\} \) to an infinite number of servers and are served immediately upon arrival. We denote by \( Q^N_\infty \) the steady-state number of jobs in the \( N \)th infinite-server queue.

Next, we provide a mathematical definition of the SITA policy prescription. Let \( \mathbb{N} \) denote the set of natural numbers, \( \mathbb{R}_+ = [0, \infty) \), \( \mathbb{Z}_+ = \mathbb{N} \cup \{0\} \). The SITA policy prescription; say, denoted by \( \pi(N) \), is fully characterized by the following parameters:

- a positive integer-valued function \( m(N) \),
- an \((m(N) + 1)\)-dimensional function \( \{r_i(N), i = 0, \ldots, m(N)\} \in \mathbb{R}_+^{m(N)+1} \), which is increasing in \( i \), and
- an \((m(N))\)-dimensional function \( \{s_i(N), i = 1, \ldots, m(N)\} \in \mathbb{Z}_+^{m(N)} \) satisfying \( \sum_{i=1}^{m(N)} s_i(N) = N \).

Specifically, in an \( N \)-server system under policy \( \pi(N) \), jobs are divided into \( m(N) \) types according to their sizes: all jobs with their size in the interval \((r_{i-1}(N), r_i(N)]\) are type \( i \) jobs, with \( r_0(N) \equiv 0 \) and \( r_m(N) \equiv N \) equal to the largest possible value of the job size (which may equal infinity). \( s_i(N) \) of the \( N \) servers form a pool processing type \( i \) jobs exclusively on an FCFS basis.

To state our main results, we start by defining two notions of optimality with respect to the rate at which \( \mathbb{P}[Q^N_\infty \geq N] \) decays to zero in the QD regime, compared to the benchmark \( \mathbb{P}[Q^N_\infty \geq N] \).

**Definition 1.** A policy \( \pi(N) \) is strongly optimal in the QD regime, if the steady-state total number of jobs in the system under \( \pi(N) \), denoted by \( Q^N_\infty \), satisfies

\[
\lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}[Q^N_\infty \geq N] = \lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}[Q^N_\infty \geq N].
\]  

**Definition 2.** A family of policies \( \{\pi_\epsilon(N)\}_{\epsilon > 0} \) are weakly optimal in the QD regime, if for all \( \epsilon > 0 \), the steady-state total number of jobs in the system under \( \pi_\epsilon(N) \), denoted by \( Q^N_\infty \), satisfies

\[
\lim sup_{N \to \infty} \frac{1}{N} \log \mathbb{P}[Q^N_\infty \geq N] - \lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}[Q^N_\infty \geq N] < \epsilon.
\]  

These two definitions make sense because, as we shall show in §3.1,

\[
\lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}[Q^N_\infty \geq N] \in (-\infty, 0)
\]  

In the next section, we formulate our model, provide the organization of the remainder of this paper, and then state our main results in more detail.
This policy is load balancing, as it follows from (6) and (7) that the traffic intensity to each server pool in the
probability \( p \) weak optimality.

Our first main result states that, if the job size \( S \) has a finite support; say, \( m \) possible values, the policy of
dividing servers into \( m \) pools each of which specializes in processing jobs with one of the \( m \) possible sizes is
strongly optimal, if the work loads are balanced among the \( m \) server pools or each pool has (asymptotically)
the same traffic intensity as that for the whole system, i.e., \( \rho \). Specifically, suppose each job size equals \( d_i \) with
probability \( p_i, i = 1, \ldots, m \), where \( \sum_{i=1}^m p_i = 1 \), \( \sum_{i=1}^m p_i d_i = \mu^{-1} \), \( 2 \leq m < \infty \), and, without loss of generality, \( d_1 < d_2 < \cdots < d_{m-1} < d_m \). We define the SITA policy \( \pi(N) \) as follows:

\[
m(N) := m, \quad r_0(N) := 0, \quad r_i(N) := d_i, \quad i = 1, \ldots, m,
\]
\[
s_i(N) := \lfloor N p_i d_i \mu \rfloor, \quad i = 1, \ldots, m-1, \quad s_m(N) = N - \sum_{i=1}^{m-1} s_i(N).
\]

This policy is load balancing, as it follows from (6) and (7) that the traffic intensity to each server pool in the
\( N \)th system; namely, \( \rho_i(N) \), reads

\[
\rho_i(N) = \rho \frac{N p_i d_i \mu}{[N p_i d_i \mu]} \approx \rho \quad \text{for} \quad i = 1, \ldots, m-1,
\]
\[
\rho_m(N) = \rho \frac{N - \sum_{i=1}^{m-1} N p_i d_i \mu}{N - \sum_{i=1}^{m-1} [N p_i d_i \mu]} \approx \rho.
\]

This load-balancing SITA policy achieves strong optimality in the QD regime.

Theorem 1. Suppose the job size \( S \) has a finite support. The SITA policy \( \pi(N) \), defined by (6) and (7),
is strongly optimal in the QD regime.

This result can be extended to systems with hyperexponentially distributed \( S \) with \( m \) branches, which are
appropriate models for many kinds of service systems, where \( m \) types of services are provided and, conditioning
on the type of service that a customer requests, his service time is exponentially distributed. Assuming that
the type of service requested by each customer is known to the system scheduler (achieved in call centers, for
example, by asking customers to select the type of service they need before assigning them to agents), a similar
policy of forming \( m \) load-balanced server pools each of which provides one type of service is strongly optimal
(see Remark 1).

Our second main result is an extension of the first result to systems with general job size distributions (see
Algorithms 1–2, Theorems 2–3, Corollary 1). Specifically, we provide a family of SITA policies that are weakly
optimal for general job size distributions. Our only assumption on \( S \) is that \( \mathbb{E}[S^\alpha] < \infty \) for some \( \alpha > 1 \), and
therefore our result holds not only for light-tailed job sizes but also for heavy-tailed ones. Although the intuition
of dividing servers into load-balanced pools remains the same with general \( S \), the prescription of the SITA
policies is much more involved; in fact, because the job size has an unbounded support while only finitely many
servers are available, all the policy parameters need to be chosen very carefully. This is the key to achieving
weak optimality.
Roughly speaking, in the general job size case, our proposed policies assign one server to process jobs whose size is above a very large threshold value, and enforce load-balancing SITA among all other servers and jobs. This is far from trivial, both when it comes to the formulation of the policies and to the associated analysis, as the number of size intervals we consider grows with the system size $N$. We next describe the policies in detail and state the optimality results.

First, we assume that $S$ is integer valued. The approach that we take in this scenario captures the essence of how weak optimality is achieved in the general unbounded support case, and we believe that both the policy construction and the proof technique used here may be useful in other contexts. Specifically, suppose that the job size $S$ can take on any $i \in \mathbb{N}$ and $E[S^\alpha] < \infty$ for some $\alpha > 1$. Define $p_i := \mathbb{P}[S = i], i \in \mathbb{N}$. Throughout the paper, we denote the floor and ceiling functions by $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$, respectively, the natural logarithm by $\log$ and define

$$I(\rho) := -\log \rho - 1 + \rho.$$  \hfill (10)

We shall see that $I(\rho)$ is the decay exponent for $\mathbb{P}\{Q_{\infty}^N \geq N\}$ (see (23)). For any given $\epsilon > 0$, we propose the SITA policy $\pi_{\epsilon}(N)$ as follows.

**Algorithm 1** ($\pi_{\epsilon}(N)$: SITA for integer-valued $S$ with finite $\alpha$th moment for $\alpha > 1$).

1. Fix $\eta \in (0, 1)$ and
   $$\gamma \in (0, \min\{\eta \alpha^{-1}(\alpha - 1)^2, 1\}).$$  \hfill (11)

2. Find $\sigma(\epsilon) \in (0, 1 - \rho)$ such that $\rho_\epsilon := \rho[1 - \sigma(\epsilon)]^{-1}$ satisfies
   $$-I(\rho) \leq -I(\rho_\epsilon) < -I(\rho) + \epsilon.$$  \hfill (12)

3. Let
   $$f_i(N) := i \quad \text{for} \quad i = 1, \ldots, \lfloor N^\eta \rfloor,$$
   $$f_i(N) := f^\alpha_{i-1}(N)N^{-\gamma - \lceil N^\eta \rceil \gamma} \quad \text{for} \quad i = \lfloor N^\eta \rfloor + 1, \ldots, 2\lfloor N^\eta \rfloor - 1.$$  \hfill (13, 14)

4. Define $\pi_{\epsilon}(N)$ as follows:
   $$m(N) := 2\lfloor N^\eta \rfloor, \quad r_0(N) := 0, \quad r_2(N) := +\infty,$$
   $$r_i(N) := \lceil f_i(N) \rceil \quad \text{for} \quad i = 1, \ldots, 2\lfloor N^\eta \rfloor - 1,$$
   $$s_i(N) := \lfloor Np_i f_i(N)(1 - \sigma(\epsilon)) \rfloor \quad \text{for} \quad i = 1, \ldots, 2\lfloor N^\eta \rfloor - 2,$$  \hfill (15, 16, 17)
   where $P_i := \mathbb{P}\{S \in (r_{i-1}(N), r_i(N)]\} = \mathbb{P}\{S \in (f_{i-1}(N), f_i(N)]\}$,
   $$s_{2\lfloor N^\eta \rfloor - 1}(N) := N - 1 - \sum_{i=1}^{2\lfloor N^\eta \rfloor - 2} s_i(N), \quad s_{2\lfloor N^\eta \rfloor}(N) := 1.$$  \hfill (18)

The recursive definition (14) is key in the policy prescription. It is significant that (14) is parameterized by the moment index of the job size distribution $\alpha$, because this parameterization enables the proposed policy to exploit the information of the job size distribution implied by the moment index.

From (13), (16), and (17), for $i = 1, \ldots, \lfloor N^\eta \rfloor$, we easily obtain that

$$r_i(N) = f_i(N) = i \quad \text{and} \quad s_i(N) = \lfloor Np_i \mu(1 - \sigma(\epsilon)) \rfloor.$$  \hfill (19)

A simple calculation using (19) further shows that the traffic intensities to the first $\lfloor N^\eta \rfloor$ server pools are all $\rho_\epsilon$ except for a round-off error, just like (7) leading to (8). Hence, under policy $\pi_{\epsilon}(N)$, the first $\lfloor N^\eta \rfloor$ subsystems are $M/D/s_i(N)$ queues, all with traffic intensity $\rho_\epsilon$.

For all $i = \lfloor N^\eta \rfloor + 1, \ldots, 2\lfloor N^\eta \rfloor - 1$, we first note that the range of the job sizes in subsystem $i$ is $(r_{i-1}(N), r_i(N)]$ by the definition of SITA; in particular, the maximum possible job size is $r_i(N)$, which is less than or equal to $f_i(N)$ according to (16). In addition, the definition of $s_i(N)$ (17) is exactly specified in such a way that if all jobs in subsystem $i$ had size $f_i(N)$, then this subsystem would also be an $M/D/s_i(N)$ queue with traffic intensity $\rho_\epsilon$ (except for a rounding error). In §5, we shall analyze the performance of these $M/D/s_i(N)$ queues, which serves as an upper bound for the original system.

Finally, (18) states that the last subsystem consists of only one server, reserved for jobs whose sizes are greater than $r_{2\lfloor N^\eta \rfloor - 1}(N)$. The traffic intensity in this single-server queue turns out to be exponentially small as $N \to \infty$ (see (114)), and therefore the steady-state number of jobs in this single-server queue is, loosely speaking, negligible in the QD regime.

In §5, we shall show that the family of policies $\{\pi_{\epsilon}(N)\}_{\epsilon > 0}$ achieve weak optimality.
THEOREM 2. Let the job size $S$ be a discrete random variable taking on positive integer values, with $p_i := \mathbb{P}\{S = i\}, i \in \mathbb{N}$, and assume $\mathbb{E}[S^\alpha] < \infty$ for some $\alpha > 1$. Then, the family of policies $\{\pi_s(N)\}_{s > 0}$ prescribed by Algorithm 1, are weakly optimal in the QD regime.

As a consequence of Theorem 2, if all values in the range of $S$ are divisible by some $\delta > 0$, an easy modification of $\{\pi_s(N)\}_{s > 0}$ (by measuring time in units of $\delta$’s, and thus viewing $S$ as integer valued) achieves weak optimality. For the sake of brevity, we omit the detailed description of the modified weakly optimal policies. We simply denote by $\{\pi_{s/\delta}(N, S)\}_{s > 0}$ the family of SITA policies that are weakly optimal for systems with job size $S$, where all possible values of $S$ are divisible by $\delta$. Note that by this definition, $\pi_s(N) = \pi_{s/\delta}(N, S)$.

COROLLARY 1. Let the job size $S$ be a discrete random variable whose possible values are integer multiples of $\delta$ for some $\delta > 0$, and $\mathbb{E}[S^\alpha] < \infty$ for some $\alpha > 1$. The family of policies $\{\pi_{s/\delta}(N, S)\}_{s > 0}$ are weakly optimal in the QD regime.

Finally, we describe the SITA policies for general job size distributions, and state the weak optimality result as Theorem 3.

Algorithm 2 ($\pi_{s/\delta}(N)$: SITA for general $S$ with finite $\alpha$th moment for $\alpha > 1$).

1. Find $\delta_0 = \delta_0(\epsilon) > 0$ such that, with

\[
S_{\delta_0} := \delta_0 \left\lceil \frac{S}{\delta_0} \right\rceil \quad \text{and} \quad \rho_{\delta_0} := \lambda \cdot \mathbb{E}[S_{\delta_0}], \tag{20}
\]

the following holds:

\[-I(\rho) \preceq -I(\rho_{\delta_0}) < -I(\rho) + \frac{1}{2}\epsilon. \tag{21}\]

2. Let $\pi_{s/\delta}(N) := \pi_{s/\delta}(N, S_{\delta_0})$ as given in Corollary 1.

THEOREM 3. Suppose that there exists $\alpha > 1$ such that $\mathbb{E}[S^\alpha] < \infty$. The family of policies $\{\pi_{s/\delta}(N)\}_{s > 0}$ prescribed by Algorithm 2, are weakly optimal in the QD regime.

3. Preliminaries.

In this section, we develop some preliminary results, which will be useful in proving our main results. These preliminary results may also be of independent interest.

3.1. Exact asymptotics for the M/G/$\infty$ queue. This subsection focuses on the exact asymptotic result on $\mathbb{P}\{Q_N^\infty \geq N\}$. For any two real sequences $(a^N)$ and $(b^N)$, we write $a^N \sim b^N$ if $\lim_{N \to \infty} (a^N/b^N) = 1$.

THEOREM 4.

\[
\mathbb{P}\{Q_N^\infty \geq N\} \sim \frac{1}{(1 - \rho)\sqrt{2\pi N}} \cdot e^{-N\rho} = N^\infty \tag{22}
\]

Therefore

\[
\lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}\{Q_N^\infty \geq N\} = -I(\rho). \tag{23}
\]

PROOF. First, because $Q_N^\infty$ is Poisson distributed with mean $N\rho$ (see Gross and Harris [13, (2.40)]), we have that, for all $i \in \mathbb{Z}_+$,

\[
\mathbb{P}\{Q_N^\infty = N + i\} = \frac{(N\rho)^i i!}{(N + i)!} \cdot \mathbb{P}\{Q_N^\infty = N\}. \tag{24}
\]

It then follows that, for all $i \in \mathbb{Z}_+$,

\[
\left( \frac{N\rho}{N + i} \right)^i \cdot \mathbb{P}\{Q_N^\infty = N\} \leq \mathbb{P}\{Q_N^\infty = N + i\} \leq \rho^i \cdot \mathbb{P}\{Q_N^\infty = N\}, \tag{25}
\]

and thus

\[
\sum_{i=0}^{\infty} \left( \frac{N\rho}{N + i} \right)^i \cdot \mathbb{P}\{Q_N^\infty = N\} \leq \mathbb{P}\{Q_N^\infty \geq N\} \leq \sum_{i=0}^{\infty} \rho^i \cdot \mathbb{P}\{Q_N^\infty = N\}. \tag{26}
\]

Applying the dominated convergence theorem to the left-hand side of the first inequality in (26) when letting $N \to \infty$, we obtain that

\[
\mathbb{P}\{Q_N^\infty \geq N\} \sim \frac{1}{1 - \rho} \cdot \mathbb{P}\{Q_N^\infty = N\}. \tag{27}
\]

The exact asymptotic result (22) follows from (27), $Q_N^\infty$ being Poisson with mean $\rho N$, and the Stirling’s approximation for the factorial. The logarithmic asymptotic result (23) then follows immediately from (22). \qed
3.2. Exact asymptotics for the M/D/N queue. In this subsection, we consider systems with deterministic job sizes equal to $\mu^{-1}$. The exact asymptotic result that we develop here is new and will be used in §4 to analyze systems where the job size has a finite support.

**Theorem 5.** Consider the M/D/N queue under FCFS in the QD regime. Let $Q^N$ denote the steady-state total number of jobs in the system. Then

$$\mathbb{P}[Q^N \geq N] \sim \mathbb{P}[Q^N_\infty \geq N].$$

**Proof.** The desired result (28) is equivalent to

$$\mathbb{P}[W^N > 0] \sim \mathbb{P}[Q^N_\infty \geq N],$$

where $W^N$ denotes the steady-state delay (or waiting time) in the $N$th system. In what follows, we prove relation (29).

Because the steady-state delay in the M/D/N queue under FCFS has the same distribution as that in the same system under cyclic scheduling (i.e., the policy under which every $N$th customer is assigned to the same server; see Jelenković et al. [18, Lemma 2]), we consider the M/D/N queue under cyclic scheduling instead, and restrict our attention to an arbitrarily chosen server in the M/D/N queue and jobs processed by this server. This subsystem is just a G/D/1 queue, where each interarrival time $X_i$, $i \geq 1$, is the sum of $N$ independent copies of $T^N_i$’s (i.e., Gamma ($N, \lambda^N$)) and each job size is $\mu^{-1}$. Because this G/D/1 queue has the same steady-state delay distribution as the whole M/D/N queue under cyclic scheduling, we shall analyze this subsystem and also denote its steady-state delay by $W^N$.

First, a fundamental result for the GI/GI/1 queue (see Asmussen [1, Chapter X, Proposition 1.1]) states that

$$W^N \overset{d}{=} \max_{k \geq 0} S_k,$$

(30)

where $\overset{d}{=}$ means “equal in distribution to,” $S_0 = 0$ and $S_k = \sum_{i=1}^{k} (\mu^{-1} - X_i)$, $k \geq 1$. Therefore we have that

$$\mathbb{P}[S_1 > 0] \leq \mathbb{P}[W^N > 0] \leq \sum_{k=1}^{\infty} \mathbb{P}[S_k > 0].$$

(31)

Also, note that, for any $k \geq 1$,

$$\mathbb{P}[S_k > 0] = \mathbb{P}\left\{ \sum_{i=1}^{k} X_i \leq k\mu^{-1} \right\} = \mathbb{P}\left\{ \sum_{i=1}^{kN} T^N_i \leq k\mu^{-1} \right\}
= \mathbb{P}[A^N(k\mu^{-1}) \geq kN] = \mathbb{P}[B^{kN} \geq kN],$$

(32)

where $B^{kN}$ denotes a Poisson random variable with mean $kN\rho$. The rest of the proof consists of the derivation of asymptotic lower and upper bounds that coincide in the limit.

**Lower Bound:** Combining $Q^N_\infty \overset{d}{=} B^N$, the first inequality in (31) and (32) with $k = 1$, we immediately have

$$\liminf_{N \to \infty} \frac{\mathbb{P}[W^N > 0]}{\mathbb{P}[Q^N_\infty \geq N]} \geq 1.$$

(33)

**Upper Bound:** It follows from the second inequality in (31) and (32) that

$$\mathbb{P}[W^N > 0] \leq \mathbb{P}[B^N \geq N] + \sum_{k=2}^{\infty} \mathbb{P}[B^{kN} \geq kN]
\leq \mathbb{P}[B^N \geq N] + \sum_{k=2}^{\infty} e^{-kN(\theta - \rho e^\theta + \rho)}$$

(34)

for all $\theta > 0$, where the last inequality is due to the Chernoff bound. In particular, letting $\theta = -\log \rho$, which is the minimizer of $-(\theta - \rho e^\theta + \rho)$, yields the following upper bound:

$$\mathbb{P}[W^N > 0] \leq \mathbb{P}[B^N \geq N] + \sum_{k=2}^{\infty} e^{-kN(\rho)}$$

(35)

$$= \mathbb{P}[B^N \geq N] + \frac{e^{-2N\rho}}{1 - e^{-N\rho}}$$

$$\leq \mathbb{P}[B^N \geq N] + \frac{e^{-2N\rho}}{1 - e^{-\rho}}.$$
Denoting the constant \([1 - e^{-N\rho}]^{-1}\) by \(C_1\) and dividing both sides of (36) by \(P\{Q_\infty^N \geq N\}\) (or equivalently, \(P\{B^N \geq N\}\)) then yields
\[
\frac{P\{W^N > 0\}}{P\{Q_\infty^N \geq N\}} \leq 1 + C_1 e^{-2N\rho}.
\] (37)

We next apply (22) to \(P\{Q_\infty^N \geq N\}\) on the right-hand side of (37), which gives that
\[
\frac{P\{W^N > 0\}}{P\{Q_\infty^N \geq N\}} \leq 1 + C_1 (1 - \rho) \sqrt{2\pi N} \cdot e^{-N\rho}[1 + o(1)],
\] (38)

where \(a^N = o(1)\) if \(\lim_{N \to \infty} a^N = 0\). Therefore
\[
\limsup_{N \to \infty} \frac{P\{W^N > 0\}}{P\{Q_\infty^N \geq N\}} \leq 1.
\] (39)

Finally, combining (33) and (39) completes the proof. □

3.3. Random walk estimates and bounds for the M/D/c queue. In this subsection, we provide a random walk result and two bound results for the M/D/c queue, which will be relied upon in the proof of our main results. Because we shall apply these results to subsystems with a different number of servers (i.e., \(s_i(N)\), \(i = 1, \ldots, m(N) - 1\)), we denote the number of servers by \(c\), instead of \(N\), in stating them.

First, we state a result concerning the maximum of a random walk process. Its proof is deferred to the end of this subsection.

**Proposition 1.** Let \(B_i, i \in \mathbb{N}\) be a sequence of i.i.d. Poisson random variables with mean \(pc\), where \(c \in \mathbb{N}\) and \(p \in (0, 1)\). Let \(M = \max_{n \geq 0} S_n\), with \(S_0 := 0\) and \(S_n := \sum_{i=1}^{n} (B_i - c)\) for \(n \geq 1\). Then, for any \(c \in \mathbb{N}\),
\[
P\{M \geq j\} \leq K(\rho) p^j \cdot P\{M \geq 1\} \quad \text{for all } j \in \mathbb{N},
\] (40)

where \(K(\rho)\) is a function of \(\rho\) only and, in particular, \(K(\rho)\) is independent of \(c\).

Now, using Proposition 1, we prove some bounds on the steady-state distribution of the total number of jobs in an M/D/c queue.

**Lemma 1.** Consider an M/D/c queue with the traffic intensity \(\rho \in (0, 1)\). Let \(Q\) denote the steady-state total number of jobs in the system. Then, for all \(j \in \mathbb{N}\),
\[
P\{Q = c - j\} \leq \rho^{-j} \cdot P\{Q = c\},
\] (41)
\[
P\{Q \geq c + j\} \leq K(\rho) p^j \cdot P\{Q \geq c\},
\] (42)

where \(K(\rho)\) is a function of \(\rho\) only.

**Proof.** In the M/D/c queue, the following relation holds (see Tijms [28, pp. 288–289])
\[
Q \overset{\text{d}}{=} (Q - c)^+ + B,
\] (43)

where \(B\) is Poisson distributed with mean \(pc\). Specifically, for any \(n \geq 0\),
\[
P\{Q = n\} = e^{-pc}(pc)^n\frac{1}{n!} \cdot \sum_{k=0}^{n} P\{Q = k\} + \sum_{k=n+1}^{\infty} P\{Q = k\} \cdot e^{-pc}(pc)^{n-k+c}\frac{1}{(n-k+c)!}.
\] (44)

We first verify (41). Substituting \(n\) in (44) with \(c\) and \(c - j\), respectively, yields that
\[
P\{Q = c\} = e^{-pc}(pc)^c\frac{1}{c!} \cdot \sum_{k=0}^{c} P\{Q = k\} + \sum_{k=c+1}^{\infty} P\{Q = k\} \cdot e^{-pc}(pc)^{2c-k}\frac{1}{(2c-k)!},
\] (45)

and for all \(j = 1, 2, \ldots, c\),
\[
P\{Q = c - j\} = e^{-pc}(pc)^{c-j}\frac{1}{(c-j)!} \cdot \sum_{k=0}^{c} P\{Q = k\} + \sum_{k=c+1}^{\infty} P\{Q = k\} \cdot e^{-pc}(pc)^{2c-j-k}\frac{1}{(2c-j-k)!}.
\] (46)

\(^1\) We suppress the dependence of \(B_i\)'s, \(S_i\)'s, and \(M\) on \(c\) in the notation.
Now, define
\[
A := \frac{e^{-pc}(pc)^c}{c!} \cdot \sum_{k=0}^{c} \mathbb{P}(Q = k), \\
B_k := \mathbb{P}(Q = k) \cdot \frac{e^{-pc}(pc)^{2c-k}}{(2c-k)!} \quad \text{for all } k = c + 1, \ldots, 2c, \\
C := \frac{e^{-pc}(pc)^{c-j}}{(c-j)!} \cdot \sum_{k=0}^{c} \mathbb{P}(Q = k), \\
D_k := \mathbb{P}(Q = k) \cdot \frac{e^{-pc}(pc)^{2c-j-k}}{(2c-j-k)!}.
\]

Substituting these definitions into (45) and (46), respectively, then yields
\[
\mathbb{P}(Q = c) = A + \sum_{k=c+1}^{2c} B_k \geq A + \sum_{k=c+1}^{2c-j} B_k \quad \text{for all } j = 1, 2, \ldots, c,
\]
and
\[
\mathbb{P}(Q = c - j) = C + \sum_{k=c+1}^{2c-j} D_k.
\]

Because \( A/C \geq \rho^j \) and \( B_k/D_k \geq \rho^j \) for all \( k = c + 1, \ldots, 2c - j \), (41) follows from dividing (49) by (50).

Next, we turn to proving (42). From (43), we have
\[
(Q - c)^+ \leq [(Q - c)^+ + B - c]^+,
\]
or
\[
Q_q \leq (Q_q + B - c)^+,
\]
where \( Q_q \) denotes the steady-state number of jobs waiting in the queue. We further let \( M = \max_{q \geq 0} S_q \), where \( S_0 = 0, S_q = \sum_{n=1}^{q} (B_n - c) \) for \( q \geq 1 \), and \( B_n \)'s are independent random variables equal in distribution to \( B \). By the standard Lindley recursion result (see Asmussen [1, Corollary 6.6, p. 94]), we have \( Q_q \leq M \). Then, we apply Proposition 1 to \( Q_q \) and arrive at
\[
\mathbb{P}(Q_q \geq j) \leq K(\rho)\rho^j \cdot \mathbb{P}(Q_q \geq 1) \quad \text{for all } j \in \mathbb{N},
\]
or equivalently,
\[
\mathbb{P}(Q \geq c + j) \leq K(\rho)\rho^j \cdot \mathbb{P}(Q_q \geq 1) \quad \text{for all } j \in \mathbb{N}.
\]
Finally, (42) follows from (53) because \( \mathbb{P}(Q_q \geq 1) > \mathbb{P}(Q \geq c) \). \( \square \)

Our next preparatory result is an exponential upper bound on \( \mathbb{P}(Q \geq c) \) in the M/D/c queue. This bound is not used in the proof for systems where the job size has a finite support, but is needed in our proof for general (discrete) job size distributions.

**Lemma 2.** Consider an M/D/c queue with the traffic intensity \( \rho \in (0, 1) \). Let \( Q \) denote the steady-state total number of jobs in the system. Then
\[
\mathbb{P}(Q \geq c) \leq \frac{e^{-cI(\rho)}}{1 - e^{-cI(\rho)}}.
\]

**Proof.** The proof is similar to the derivation of (35) from (34). Here, in addition to applying the minimizing Chernoff bound to the summation from \( k = 2 \) to \( \infty \), we apply it to the first term of (34) as well. Therefore, with \( N \) in (35) replaced by \( c \), we obtain that
\[
\mathbb{P}(Q \geq c) \leq \sum_{k=1}^{\infty} e^{-kcI(\rho)} = \frac{e^{-cI(\rho)}}{1 - e^{-cI(\rho)}} \leq \frac{e^{-cI(\rho)}}{1 - e^{-cI(\rho)}}. \quad \square
\]

Finally, we prove the random walk estimate result. Our proof is related to the light traffic analysis of random walks (see Asmussen [1]).

**Proof of Proposition 1.**

**Step 1.** We prove an inequality concerning the increment of the random walk (i.e., \( B_1 - c \)): for any \( c \in \mathbb{N} \),
\[
\mathbb{P}(B_1 - c \geq j) \leq K_1(\rho)\rho^j \cdot \mathbb{P}(B_1 - c \geq 1) \quad \text{for all } j \in \mathbb{N},
\]
where \( K_1(\rho) \) is a function of \( \rho \) only.
From (25) (with $Q^N$ and $N$ replaced by $B_i$ and $c$, respectively), we obtain that

$$\mathbb{P}(B_i - c \geq j) \leq \frac{\rho^j}{1 - \rho} \cdot \mathbb{P}(B_i = c) \quad \text{for all } j \in \mathbb{N},$$

and

$$\mathbb{P}(B_i - c \geq 1) \geq \sum_{i=1}^{\infty} \left( \frac{c \rho}{c + i} \right)^i \cdot \mathbb{P}(B_i = c) \geq \sum_{i=1}^{\infty} \left( \frac{\rho}{1 + i} \right)^i \cdot \mathbb{P}(B_i = c) \geq \frac{\rho}{2} \cdot \mathbb{P}(B_i = c).$$

Then, dividing (57) by (58) yields (56), where $K_i(\rho) = 2(1 - \rho)\rho^{-i}$.

**Step 2.** We prove a similar bound concerning the first ascending ladder height of the random walk; namely, $S_{\tau_i}$, where $\tau_i = \inf\{n \geq 1: S_n > 0\}$. Specifically, we show that, for any $c \in \mathbb{N}$,

$$\mathbb{P}(S_{\tau_i} \geq j) \leq K_2(\rho) \rho^j \cdot \mathbb{P}(S_{\tau_i} \geq 1) \quad \text{for all } j \in \mathbb{N},$$

where $K_2(\rho)$ is a function of $\rho$ only. Note that $S_{\tau_i}$ is defined as 0 when $\tau_i = \infty$.

Following the same notation as used in Asmussen [1, pp. 221–223], we define $\tau_- := \inf\{n \geq 1: S_n \leq 0\}$ and $\tau_- (i + 1) := \inf\{n > \tau_- (i): S_n \leq S_{\tau_- (i)}\}$, where $\tau_- (1) := \tau_-$. Note that, the descending ladder heights are not strict; in particular, unlike $S_{\tau_i}$, $S_\tau$ can be 0, even when $\tau_- < \infty$.

First, we apply Asmussen [1, Equation (1.7), p. 269], with the set $A$ in that expression replaced by $[j, \infty)$, and obtain that

$$\mathbb{P}(S_{\tau_i} \geq j) = \sum_{x=-\infty}^{0} \mathbb{P}(B_i - c \geq j - x) \cdot \left[ I\{x = 0\} + \sum_{i=1}^{\infty} \mathbb{P}(S_{\tau_i} (i) = x) \right] = \mathbb{P}(B_i - c \geq j) + \sum_{x=-\infty}^{0} \mathbb{P}(B_i - c \geq j - x) \cdot \sum_{i=1}^{\infty} \mathbb{P}(S_{\tau_i} (i) = x) = \mathbb{P}(B_i - c \geq j) + \sum_{k=0}^{\infty} R(k) \cdot \mathbb{P}(B_i - c \geq j + k),$$

where $R(k) := \sum_{i=0}^{\infty} \mathbb{P}(S_{\tau_i} (i) = -k)$ for all $k \in \mathbb{Z}^+$.

Next, we establish an upper bound of $R(k)$, which is uniform in $k \in \mathbb{Z}^+$. For all $k \in \mathbb{Z}^+$, we have

$$R(k) = \mathbb{E}[\#\{i \geq 1: S_{\tau_i} (i) = -k\}] \leq \mathbb{E}[\#\{i \geq 1: S_{\tau_i} (i) = -k\} \mid S_{\tau_i} (i) = -k \text{ for some } i \in \mathbb{N}] = 1 + \sum_{i=1}^{\infty} i \cdot \mathbb{P}(S_{\tau_i} = 0)^i \cdot \mathbb{P}(S_{\tau_i} < 0) \leq 1 + \sum_{i=1}^{\infty} i \cdot \mathbb{P}(S_{\tau_i} = 0)^i,$$

where (61) holds because $\{S_{\tau_i (i+1)} - S_{\tau_i (i)}, i \geq 1\}$ is a sequence of i.i.d. random variables equal in distribution to $S_{\tau_i}$.

Also, since $\{S_{\tau_i} = 0\} \subset \{B_i \geq c\}$, we have that

$$\mathbb{P}(S_{\tau_i} = 0) \leq \mathbb{P}(B_i \geq c) \leq e^{-c(\rho)} \leq e^{-c(\rho)} < 1,$$

where the third last inequality follows from applying the minimizing Chernoff bound (in the same way as we obtain (35)). Then, with $\delta_1(\rho) := e^{-c(\rho)}$, combining (62) and (63) leads to the following uniform bound:

$$R(k) \leq 1 + \sum_{i=1}^{\infty} i \cdot \delta_1(\rho)^i = \delta_2(\rho) \quad \text{for all } k \in \mathbb{Z}^+,$$

where $\delta_2(\rho) = 1 + \delta_1(\rho) \cdot [1 - \delta_1(\rho)]^{-2} > 1$.

---

Note that in their notation, $G_c$ corresponds to the distribution of $S_{\tau_i}$, i.e., for set $A$, $G_c(A) = \mathbb{P}(S_{\tau_i} \in A)$. Also, for set $A$, $U_c(A) := \sum_{i=0}^{\infty} G_i^c(A)$, where $G_i^c(A) = 1[0 \in A]$ and $G_i^c$, $i \geq 1$, is the $i$th convolution of $G_c$, with $G_c$ being the distribution of $S_{\tau_i}$; more specifically, $U_c(A) = 1[0 \in A] + \sum_{i=0}^{\infty} \mathbb{P}(S_{\tau_i} \in A)$, by noting that for all $i \geq 1$, $G_i^c$ is exactly the distribution of $S_{\tau_i0}$. Also, $F$ in their expression (1.7) denotes the distribution of the random walk increment, which means in our case $F(A) = \mathbb{P}(B_i - c \in A)$. 
Finally, combining (56), (60), (64), and using \(\mathbb{P}[B_1 - c \geq 1] \leq \mathbb{P}[S_{r_c} \geq 1]\) yields (59) as follows:

\[
\mathbb{P}[S_{r_c} \geq j] \leq K_1(\rho) \rho^j \cdot \mathbb{P}[B_1 - c \geq 1] + \sum_{k=0}^{\infty} \delta_2(\rho) \cdot K_1(\rho) \rho^{j+k} \cdot \mathbb{P}[B_1 - c \geq 1]
\]

\[
\leq K_1(\rho) \rho^j \cdot \mathbb{P}[S_{r_c} \geq 1] + \sum_{k=0}^{\infty} \delta_2(\rho) \cdot K_1(\rho) \rho^{j+k} \cdot \mathbb{P}[S_{r_c} \geq 1]
\]

\[
= K_2(\rho) \rho^j \cdot \mathbb{P}[S_{r_c} \geq 1],
\]

where \(K_2(\rho) = K_1(\rho) + \delta_2(\rho)K_1(\rho) \cdot (1 - \rho)^{-1}\).

Step 3. We eventually prove (40). First, there exists a constant \(\beta \in (0, \rho)\), whose value is independent of \(c\), such that

\[
\mathbb{E}[\beta^{-b_1 - \epsilon}] = 1.
\]

Specifically, using the fact that \(B_1\) is Poisson with mean \(\rho c\), a straightforward calculation shows that the desired \(\beta\) satisfies \(\rho (1 - 1/\beta) = \log \beta\).

Next, using \(\beta \in (0, \rho)\) and applying the Kolmogorov’s inequality for (sub)martingales to \(\beta^{-S_n}, n \in \mathbb{Z}_+\), we have

\[
\mathbb{P}(M \geq j) = \mathbb{P}(\beta^{-M} \geq \beta^{-j}) \leq \beta^j \text{ for all } j \in \mathbb{Z}_+.
\]

In other words, \(M \leq_{st} M_\beta\), where the subscript \(st\) stands for the usual stochastic order and \(\mathbb{P}(M_\beta = j) = (1 - \beta)\beta^j\) for all \(j \in \mathbb{Z}_+\).

Finally, using \(S_{r_c} \geq 1 = [M \geq 1]\) and \(M | S_{r_c} \geq 1 \leq G + M\), with \(G := S_{r_c} | S_{r_c} \geq 1\), we obtain that, for all \(j \in \mathbb{N}\),

\[
\mathbb{P}(M \geq j | M \geq 1) = \mathbb{P}(G + M \geq j) \leq \mathbb{P}(G + M_\beta \geq j)
\]

\[
= \sum_{k=0}^{j-1} \mathbb{P}(M_\beta = k) \cdot \mathbb{P}(G \geq j - k) + \mathbb{P}(M_\beta \geq j)
\]

\[
= \sum_{k=0}^{j-1} \mathbb{P}(M_\beta = k) \cdot \mathbb{P}(S_{r_c} \geq j - k | S_{r_c} \geq 1) + \mathbb{P}(M_\beta \geq j)
\]

\[
\leq \sum_{k=0}^{j-1} (1 - \beta)\beta^k \cdot K_2(\rho) \rho^{j+k} + \beta^j
\]

\[
< \rho^j \sum_{k=0}^{\infty} (1 - \beta) \left( \frac{\beta}{\rho} \right)^k K_2(\rho) + \beta^j
\]

\[
= K(\rho) \rho^j,
\]

where (67) follows from (59) and \(K(\rho) = K_2(\rho)(1 - \beta) \cdot (1 - \beta \rho^{-1})^{-1} + 1\). This is equivalent to (40). \(\square\)

3.4. Combinatorial results. This subsection contains two combinatorial results, which are used in the proof of our main theorems. We start by giving some necessary notation. Throughout this paper, we use the overline (or bar) symbol to denote vectors, e.g., \(\bar{c} = (c_1, c_2, \ldots, c_{m-1}, c_m)\). For any \(\bar{c} = (c_1, c_2, \ldots, c_{m-1}, c_m) \in \mathbb{Z}_+^m\), we define \(\|\bar{c}\| := \sum_{i=1}^{m} c_i\). For any \(\bar{c} \in \mathbb{Z}_+^m\) and \(j \in \mathbb{Z}_+\),

\[
S(\bar{c}, j) := \{ \bar{x} \in \mathbb{Z}_+^m : \|\bar{x}\| = \|\bar{c}\| + j \},
\]

or in words, \(S(\bar{c}, j)\) denotes the set of nonnegative integer solutions to \(\|\bar{x}\| = \|\bar{c}\| + j\). The first combinatorial result that we use in later proofs is on the cardinality of \(S(\bar{c}, j)\). From Stanley [26, p. 15], we have that

\[
|S(\bar{c}, j)| = \binom{\|\bar{c}\| + j + m - 1}{m - 1}.
\]

The second combinatorial result that we need (see Graham et al. [12, p. 1077]) is

\[
\binom{n}{k} \leq \left( \frac{ne}{k} \right)^k \text{ for all } k = 0, \ldots, n.
\]
4. Job sizes with finite support. In this section, we consider the case in which the job size $S$ has a finite support and prove Theorem 1. The following lemma is used in our proof of Theorem 1. We include its proof in the appendix.

**Lemma 3.** Consider $m$ independent $M/D/c_i$ queues, $i = 1, \ldots, m$, all with the same traffic intensity $\rho_{c_i}$. Let $Q_i$ denote the steady-state total number of jobs in the $M/D/c_i$ queue, $i = 1, \ldots, m$. Then, for all $j \in \mathbb{Z}_+$,

$$\max_{n \in \mathbb{N}} \mathbb{P}[Q_i = n, i = 1, \ldots, m] \leq K(\rho_{c_i})^n \rho_{c_i}^j \cdot \mathbb{P}[Q_i \geq c_i \text{ for all } i = 1, \ldots, m],$$

(72)

where $K(\cdot)$ is the same as given in Lemma 1.

Next, we prove Theorem 1.

**Proof of Theorem 1.** Let $Q_N$ be the steady-state total number of jobs in the $N$th system under policy $\pi(N)$. We need to prove (1). Due to (4) and (23), it suffices to show that

$$\limsup_{N \to \infty} \frac{1}{N} \log \mathbb{P}[Q^N \geq N] \leq -I(\rho).$$

(73)

First, from (8) and (9), we easily have that $\lim_{N \to \infty} \rho_{c_i}(N) = \rho$ for $i = 1, \ldots, m - 1$, and $\rho_{c_m}(N) \leq \rho$. Then, for any $\varepsilon \in (0, 1 - \rho)$, there exists an $N_\varepsilon \in \mathbb{N}$, such that, for any $N > N_\varepsilon$,

$$\rho_{c_i}(N) \leq \rho_{c_i}$$

(74)

for all $i = 1, \ldots, m$.

Let us fix $\varepsilon \in (0, 1 - \rho)$ and construct an auxiliary sequence, for which we only define the $N$th system if $N \geq N_\varepsilon$. In the $N$th system of the auxiliary sequence, there are $N$ servers and jobs arrive according to the process $\{A^N(t), t \geq 0\}$. A job is type $i$ with probability $p_i$, and type $i$ jobs have size $s_i(N) \rho_{c_i} \cdot (\lambda^N p_i)^{-1}$, $i = 1, \ldots, m$. Also, $s_i(N)$ of the $N$ servers are dedicated to processing type $i$ jobs, $i = 1, \ldots, m$, on an FCFS basis. In other words, the $N$th system of the auxiliary sequence is the same as the original $N$th system, except that the size of type $i$ jobs is $s_i(N) \rho_{c_i} \cdot (\lambda^N p_i)^{-1}$, instead of $d_i$, $i = 1, \ldots, m$.

By basic properties of the Poisson process, for systems either in the original sequence or the auxiliary sequence, the job arrivals to the $m$ server pools are independent Poisson processes with rate $\lambda^N p_i$, $i = 1, \ldots, m$. Therefore, the $m$ server pools can be viewed as $m$ subsystems operating completely independently. Furthermore, each subsystem is an $M/D/s_i(N)$ queue.

By the construction of the auxiliary sequence, all the $m$ server pools in each auxiliary system have the same traffic intensity $\rho_{c_i}$. Since all subsystems are $M/D/s_i(N)$ queues and each subsystem in the $N$th auxiliary system has the same number of servers and job arrival process as the corresponding one in the $N$th original system, (74) then implies

$$Q^N = \sum_{i=1}^m Q^N_i \leq \sum_{i=1}^m Q^N_{i, a}$$

(75)

for all $i = 1, \ldots, m$,

where $Q^N_i$ denotes the steady-state total number of type $i$ jobs in the $N$th original system and $Q^N_{i, a}$ denotes the steady-state total number of type $i$ jobs in the $N$th auxiliary system. Therefore

$$Q^N = \sum_{i=1}^m Q^N_i \leq \sum_{i=1}^m Q^N_{i, a}. $$

(76)

Define $\tilde{s}_N := (s_1(N), s_2(N), \ldots, s_{m-1}(N), s_m(N))$. For any $N > N_\varepsilon$, it follows from (76) that

$$\mathbb{P}[Q^N \geq N] = \mathbb{P}\left\{ \sum_{i=1}^m Q^N_i \geq N \right\} \leq \mathbb{P}\left\{ \sum_{i=1}^m Q^N_{i, a} \geq N \right\} = \sum_{j=0}^{\infty} \left\{ \mathbb{P}\left\{ \sum_{i=1}^m Q^N_{i, a} = N+j \right\} \right\} = \sum_{j=0}^{\infty} \sum_{n \in \mathbb{N}} \mathbb{P}[Q^N_{i, a} = n, i = 1, \ldots, m].$$

(77)

From (70), we know

$$|S(\tilde{s}_N, f)| = \frac{(N+j+m-1)!}{m-1!}.$$

(78)
Applying Lemma 3 and (78) to (77) then leads to

$$\mathbb{P}\{Q^N \geq N\} \leq \sum_{j=0}^{\infty} \left(\frac{(N+j+m-1)e^m}{m-1}\right)^{m-1} K(\rho_e)^m \rho_e^j \mathbb{P}\{Q^N_{i,u} \geq s_i(N)\} \text{ for all } i = 1, \ldots, m. \quad (79)$$

We next apply (71) to (79) and obtain that

$$\mathbb{P}\{Q^N \geq N\} \leq \sum_{j=0}^{\infty} \left(\frac{(N+j+m-1)e^m}{m-1}\right)^{m-1} K(\rho_e)^m \rho_e^j \mathbb{P}\{Q^N_{i,u} \geq s_i(N)\} \text{ for all } i = 1, \ldots, m \right) \cdot N^{m-1} \cdot \frac{1}{(m-1)^{m-1}(1-\rho_e)}[1 + o(1)], \quad (80)$$

where the last equality follows from \(\sum_{j=0}^{\infty} (N+j+m-1)^{m-1} \rho_e^j = (N^{m-1}/(1-\rho_e))[1 + o(1)]\) by the monotone convergence theorem.

From (22), (28), \(s_i(N) \sim N\rho_d\mu\), and the independence among subsystems, we have

$$\mathbb{P}\{Q^N_{i,u} \geq s_i(N)\} \text{ for all } i = 1, \ldots, m = \prod_{i=1}^{m} \mathbb{P}\{Q^N_{i,u} \geq s_i(N)\} = \frac{1}{(2\pi N)^{m/2} \sqrt{\prod_{i=1}^{m} p_i d_i \mu}} \cdot e^{-N(\rho_d)\mu}[1 + o(1)]. \quad (81)$$

Combining (80) and (81) then yields that

$$\mathbb{P}\{Q^N \geq N\} \leq \frac{K(\rho_e)^m e^{m-1} N^{(m/2)-1}}{(2\pi)^{m/2} (m-1)^{m-1}(1-\rho_e)^{m+1} \sqrt{\prod_{i=1}^{m} p_i d_i \mu}} \cdot e^{-N(\rho_d)\mu}[1 + o(1)]. \quad (82)$$

Taking logarithms, dividing both sides of (82) by \(N\), and letting \(N \to \infty\), we have

$$\limsup_{N \to \infty} \frac{1}{N} \log \mathbb{P}\{Q^N \geq N\} \leq -I(\rho_e). \quad (83)$$

Finally, letting \(\epsilon \to 0\) in (83) yields (73) and this completes the proof. \(\square\)

From Theorem 1 and its proof, it is easy to deduce that under policy \(\pi(N)\),

$$\lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}\{Q^N \geq N\} = \lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}\{Q^N \geq s_i(N)\} \text{ for all } i = 1, \ldots, m. \quad (84)$$

This suggests that, when the system operates under the load-balancing SITA, the most likely way for the event \(Q^N \geq N\) to happen is the occurrence of \(Q^N_i \geq s_i(N)\) for all \(i = 1, \ldots, m\).

**Remark 1.** For systems with \(m\) classes of customers, where class \(i\) customers’ service times are exponential with mean \(d_i\), \(i = 1, \ldots, m\), a similar load-balancing, size-based (or equivalently, class-based) task assignment policy is strongly optimal in the QD regime. This type of model and its variants, which are often proper for service systems with human customers, have been studied in other contexts; for example, see Tezcan and Dai [27] and the references therein.

Specifically, if a customer belongs to class \(i\) with probability \(p_i\), \(i = 1, \ldots, m\), the optimal policy is just to allocate \(s_i(N)\) servers for serving class \(i\) customers exclusively, where \(s_i(N)\)'s are the same as defined in (7). The strong optimality holds in this case, because most results that we have proved so far for M/D/1 queues also hold true if the service times or job sizes are exponentials.

More specifically, for the M/M/N queue, (28) can be proved as follows. With \(B^N \sim \text{Poisson}(N\rho)\), we rewrite expression (2.2) in Whitt [32] as

$$\mathbb{P}\{Q^N \geq N\} = \frac{\mathbb{P}\{B^N = N\}}{\rho \mathbb{P}\{B^N = N\} + (1-\rho) \mathbb{P}\{B^N \leq N\}} = \frac{\mathbb{P}\{B^N = N\}}{\mathbb{P}\{B^N \geq N\}} = \frac{\mathbb{P}\{B^N = N\}}{1-\rho} \cdot [1 + o(1)], \quad (85)$$
where the second last equality holds, because \( P[B^N = N] = o(1) \) and \( P[B^N \leq N] = 1 + o(1) \) by applying the weak law of large numbers to \( B^N \) (which is equal to the sum of \( N \) independent Poisson(\( \rho \)) random variables). Then, combining (27) and (85) yields (28). For the M/M/c queue, both (41) and (42) easily follow from the exact formula for the distribution of \( Q \). In fact, (42) holds as an equality with \( K(\rho) = 1 \) in that case. As a consequence, Lemma 3 also holds for \( m \) M/M/\( c \) queues with \( K(\rho_0) = 1 \); indeed, one may prove a stronger result:

\[
\max_{i \in S(i, j)} P\{Q_i = n_i, i = 1, \ldots, m\} = \rho_0^i \cdot P\{Q_i = c_i\} \text{ for all } i = 1, \ldots, m. \tag{86}
\]

Therefore the whole proof of Theorem 1 can be applied with minor changes to establish the strong optimality of the load-balancing, size-based (or class-based) task assignment policy.

5. Discrete job sizes. In this section, we prove Theorem 2, which implies Corollary 1. We start by giving an intuitive explanation of the policy construction.

With respect to the performance measure \( P[\text{steady-state total number of jobs in the system} \geq \text{total number of servers}] \), we learn from Theorem 1 that finitely many load-balanced M/D/\( \cdot \) systems perform as well as an infinite-server queue in the QD regime. In fact, in the proof of Theorem 1, one may easily verify that the step from (82) to (83) would still hold if \( m \) were replaced by some other \( o(N/\log N) \) quantity, and thus \( o(N/\log N) \) many load-balanced M/D/\( \cdot \) systems also perform as well as an infinite-server queue. This observation leads to our choice of \( m(N) \) (see (15)) at the order of \( N^\eta \) for some \( \eta \in (0, 1) \), where \( N^\eta = o(N/\log N) \).

Therefore the main idea is to construct a family of SITA policies, under which the original system consists of two components: the first component can be bounded from above (with respect to our performance measure) by \( o(N/\log N) \) many load-balanced M/D/\( \cdot \) queues, and the second component is a single-server queue, where this one server only processes jobs with extremely large sizes—specifically, growing superexponentially fast as \( N \to \infty \), cf. (110)—such that the contribution from the single-server queue to the total system size is negligible (cf. (104)). Also, note that the assumption on the finiteness of the \( \alpha \)th moment of \( S \) for \( \alpha > 1 \) provides us with a control on the tail distribution of the job size, which is critically relied upon in the policy construction and the proof.

Before analyzing the performance of the policy \( \pi_\epsilon(N) \), we first show that it is feasible for a large enough \( N \).

**Proposition 2.** For any \( \epsilon > 0 \), there exists \( N_{\epsilon,1} \in \mathbb{N} \), such that, for all \( N > N_{\epsilon,1} \), the policy \( \pi_\epsilon(N) \) prescribed by Algorithm 1 is a feasible SITA policy.

**Proof.** First, by the continuity of \( I(\cdot) \) and its monotonicity in the interval \((0, 1)\) for any \( \epsilon > 0 \), there exists \( \sigma(\epsilon) \in (0, 1 - \rho) \), such that, with \( \rho_\epsilon := \rho[1 - \sigma(\epsilon)]^{-1} \), (12) holds.

Next, we show that \( \{f_i(N), i = 1, \ldots, 2[N^\gamma]\} \) is increasing in \( i \). It is sufficient to show \( f_i(N), i = 1, \ldots, 2[N^\gamma] - 1 \) is increasing in \( i \). It easily follows from (13) that

\[
f_i(N) < \cdots < f_{2[N^\gamma]}(N). \tag{87}
\]

From the recursive definition (14), for \( i = [N^\gamma], \ldots, 2[N^\gamma] - 1 \),

\[
f_i(N) = [N^\gamma]^{\alpha - [N^\gamma]} N^{-\gamma[\alpha - [N^\gamma]] - 1 - 2\alpha - [N^\gamma] - \cdots + (i - [N^\gamma] - 1)\alpha + (i - [N^\gamma])}. \tag{88}
\]

From (14) and (88), we have for all \( i = [N^\gamma] + 1, \ldots, 2[N^\gamma] - 1 \),

\[
\frac{f_i(N)}{f_{i-1}(N)} = f_{i-1}^{-1}(N)N^{-i([N^\gamma])\gamma} \\
\geq \frac{(N\eta)^{\gamma[\alpha - [N^\gamma]] - 1 - 2\alpha - [N^\gamma] - \cdots + (i - [N^\gamma] - 1)\alpha + (i - [N^\gamma])\gamma}}{N^{i([N^\gamma])\gamma \alpha}} < N^{\eta\alpha^{i - [N^\gamma] - 1}(\alpha - 1) - \gamma((\alpha^{i - [N^\gamma]} - 1)/(\alpha - 1))}. \tag{89}
\]

Because (11) implies that

\[
\gamma < \frac{\eta(\alpha - 1)^2}{\alpha} \cdot \frac{\alpha^{i - [N^\gamma]}}{\alpha^{i - [N^\gamma] - 1} - 1} \text{ for all } i = [N^\gamma] + 1, \ldots, 2[N^\gamma] - 1, \tag{90}
\]

the exponent in (89) is positive; namely,

\[
\eta\alpha^{i - [N^\gamma] - 1}(\alpha - 1) - \gamma \frac{\alpha^{i - [N^\gamma]} - 1}{\alpha - 1} > 0, \tag{91}
\]

Therefore the whole proof of Theorem 1 can be applied with minor changes to establish the strong optimality of the load-balancing, size-based (or class-based) task assignment policy.
and thus $f_i(N)/f_{i-1}(N) > 1$, or $f_i(N)$ is increasing in $i$ for all $i = [N^\gamma], \ldots, 2[N^\gamma] - 1$. This, together with (87), establishes the assertion that $\{f_i(N), i = 1, \ldots, 2[N^\gamma] - 1\}$ is increasing in $i$.

To check feasibility, the last condition that we need to verify is that $s_{2[N^\gamma]-1}(N)$ as defined by (18) is positive. We shall show that, for any $\varepsilon > 0$, there exists $N_{\varepsilon, 1} \in \mathbb{N}$, such that, for all $N > N_{\varepsilon, 1}$,
\[
s_{2[N^\gamma]-1}(N) - \left[NP_{L_{2[N^\gamma]-1}, f_{2[N^\gamma]-1}}(N)\mu(1 - \sigma(\varepsilon))\right] > 0,
\]
(92)

or
\[
N - 1 - \sum_{i=[N^\gamma]+1}^{2[N^\gamma]-2} s_i(N) - \left[NP_{L_{2[N^\gamma]-1}, f_{2[N^\gamma]-1}}(N)\mu(1 - \sigma(\varepsilon))\right] > 0.
\]
(93)

Inequality (92) is obviously a stronger statement than $s_{2[N^\gamma]-1}(N) > 0$ and we shall need it in later proofs.

Define
\[
\Sigma_0 := \sum_{i=[N^\gamma]+1}^{2[N^\gamma]-2} s_i(N) + \left[NP_{L_{2[N^\gamma]-1}, f_{2[N^\gamma]-1}}(N)\mu(1 - \sigma(\varepsilon))\right].
\]
(94)

From (17) and (19), we have that
\[
\Sigma_0 = \sum_{i=[N^\gamma]+1}^{2[N^\gamma]-2} s_i(N) + \left[NP_{L_{2[N^\gamma]-1}, f_{2[N^\gamma]-1}}(N)\mu(1 - \sigma(\varepsilon))\right]
\]
\[
\leq \left(N(1 - \sigma(\varepsilon)) + [N^\gamma]\right) + \left(N\mu(1 - \sigma(\varepsilon))\sum_{i=[N^\gamma]+1}^{2[N^\gamma]-1} P_i f_i(N) + [N^\gamma] - 1\right)
\]
\[
= N(1 - \sigma(\varepsilon)) + 2[N^\gamma] - 1 + N\mu(1 - \sigma(\varepsilon))\sum_{i=[N^\gamma]+1}^{2[N^\gamma]-1} P_i f_i(N),
\]
(95)

where the terms $[N^\gamma]$ and $[N^\gamma] - 1$ in (95) count the maximum possible rounding error. Furthermore,
\[
\sum_{i=[N^\gamma]+1}^{2[N^\gamma]-1} P_i f_i(N) \leq E[S^\alpha] \cdot \sum_{i=[N^\gamma]+1}^{2[N^\gamma]-1} f_{i-1}^\alpha f_i(N),
\]
(97)

because by the Markov inequality,
\[
P_i \geq \mathbb{P}[\{S \in (f_{i-1}(N), f_i(N))\}] < \mathbb{P}[S > f_{i-1}(N)] \leq E[S^\alpha] \cdot f_{i-1}^\alpha(N).
\]
(98)

Also, from (14), we have
\[
\sum_{i=[N^\gamma]+1}^{2[N^\gamma]-1} f_{i-1}^\gamma f_i(N) = \sum_{i=[N^\gamma]+1}^{2[N^\gamma]-1} N^{-i-[N^\gamma]} \leq \sum_{i=1}^{\infty} N^{-i\gamma}.
\]
(99)

By applying (97) and (99) to (96), we obtain that
\[
S_0 \leq N(1 - \sigma(\varepsilon)) + 2[N^\gamma] - 1 + N\mu(1 - \sigma(\varepsilon))E[S^\alpha] \cdot \sum_{i=1}^{\infty} N^{-i\gamma}
\]
\[
= N(1 - \sigma(\varepsilon)) + o(N) < N - 1
\]
(100)

for any $N > N_{\varepsilon, 1}$, where $N_{\varepsilon, 1}$ is some positive integer depending on $\varepsilon$ (through $\sigma(\varepsilon)$) only. This is equivalent to (93) and completes the feasibility proof. □

The following lemma is needed in our proof of Theorem 2. We postpone its proof until the appendix.

**Lemma 4.** For any $\eta \in (0, 1)$ and $\rho_\varepsilon \in (0, 1)$,
\[
\sum_{j=0}^{\infty} \left(\frac{N + 2[N^\gamma] + j - 3}{2[N^\gamma] - 2}\right)^{2[N^\gamma]-2} \rho_\varepsilon \leq \frac{(\varepsilon N)^{2[N^\gamma]-2}}{1 - \rho_\varepsilon}[1 + o(1)].
\]
(101)

Finally, we provide the proof of Theorem 2.
Proof of Theorem 2. Fix any $\epsilon > 0$. Let the sequence of systems be under policy $\pi_\epsilon(N)$ and we consider $N > N_{\epsilon,1}$ as specified by Proposition 2. Let $Q^N_{i,e}, i = 1, \ldots, 2[N^\eta]$ be the steady-state total number of type $i$ jobs in the $N$th system, and

$$Q^N_i = \sum_{e=1}^{2[N^\eta]} Q^N_{i,e}. \quad (102)$$

Using the independence among subsystems, we have that

$$\mathbb{P}\{Q^N_i \geq N\} \leq \mathbb{P}\{Q^N_{2[N^\eta],i} \geq 1\} + \mathbb{P}\left\{\sum_{i=1}^{2[N^\eta]-1} Q^N_{i,e} \geq N - 1\right\}. \quad (103)$$

Therefore, if we can prove that

$$\mathbb{P}\{Q^N_{2[N^\eta],i} \geq 1\} \leq e^{-N\rho_\epsilon}, \quad (104)$$

and

$$\limsup_{N \to \infty} \frac{1}{N} \log \mathbb{P}\{Q^N_i \geq N\} \leq -I(\rho_\epsilon). \quad (105)$$

then it follows from the principle of the largest term (see Ganesh et al. [10], Lemma 2.2) that

$$\limsup_{N \to \infty} \frac{1}{N} \log \mathbb{P}\{Q^N_i \geq N\} \leq -I(\rho_\epsilon). \quad (106)$$

This, together with (12), would establish the weak optimality. In what follows, we shall show (104) and (105).

First, we prove that there exists some $N_{\epsilon,2} \in \mathbb{N}$, whose value only depends on $\epsilon$, such that, for any $N > N_{\epsilon,2}$, (104) holds. Replacing $i$ by $2[N^\eta] - 1$ in (88) yields

$$f_{2[N^\eta]-1}(N) = [N^\eta]\alpha^{[N^\eta]-1}N^{-\gamma(\alpha [N^\eta] - 2 + 2a[N^\eta] - 3) + \cdots + ([N^\eta] - 2)\alpha + ([N^\eta] - 1)}$$

$$\geq N^{\eta\alpha^{[N^\eta]-1}[\alpha [N^\eta] - 2 + 2a[N^\eta] - 3] + \cdots + ([N^\eta] - 2)\alpha + ([N^\eta] - 1)} \quad (107)$$

Using $\alpha > 1$, we obtain the following lower bound for the exponent in (107):

$$\eta\alpha^{[N^\eta]-1} - \gamma(\alpha [N^\eta] - 2 + 2a[N^\eta] - 3 + \cdots + ([N^\eta] - 2)\alpha + ([N^\eta] - 1))$$

$$= \alpha^{[N^\eta]}(\eta - \gamma(\alpha - 2a - 3 - \cdots - ([N^\eta] - 1)\alpha))$$

$$\geq \alpha^{[N^\eta]}(\eta - \gamma N_{\alpha} \sum_{j=2}^{\infty} (\alpha - 1)^{-j})$$

$$= \alpha^{[N^\eta]}\left(\eta - \frac{\gamma}{(\alpha - 1)^2}\right). \quad (108)$$

For convenience, we define

$$C := \eta - \frac{\gamma}{(\alpha - 1)^2} = \frac{\eta(\alpha - 1)^2 - \gamma\alpha}{\alpha(\alpha - 1)^2}, \quad (109)$$

and it follows from (11) that $C > 0$. Then, combining (107) and (108), we have

$$f_{2[N^\eta]-1}(N) \geq N^{C_{\alpha}[N^\eta]} \geq N^{C_{\alpha}[N^\eta]}. \quad (110)$$

With $\mathbb{1}[A]$ denoting the indicator function of set $A$, we obtain the traffic intensity of the $(2[N^\eta])$-th subsystem (i.e., the single-server queue processing jobs with their size larger than $f_{2[N^\eta]-1}(N)$) as follows:

$$\rho_{2[N^\eta]}(N) = N\lambda \cdot \mathbb{E}[S \cdot \mathbb{1}[S > f_{2[N^\eta]-1}(N)] = N\lambda \cdot \mathbb{E}[S \cdot \mathbb{1}[S > f_{2[N^\eta]-1}(N)]$$

$$\leq N\lambda \cdot \mathbb{E}[\max[S^\alpha]^1/\alpha \cdot P[S > f_{2[N^\eta]-1}(N)]^{1-1/\alpha} \quad (111)$$

$$\leq N\lambda \cdot \mathbb{E}[\max[S^\alpha]^1/\alpha \cdot (\mathbb{E}[S^\alpha] f_{2[N^\eta]-1}(N)]^{\gamma \alpha})^{1-1/\alpha} \quad (112)$$

$$\leq N\lambda \cdot \mathbb{E}[\max[S^\alpha] \cdot (N^{C_{\alpha}[\alpha] - 1})^{-1} \quad (113)$$

$$= o(e^{-N\rho_\epsilon}). \quad (114)$$
where (111) is due to Hölder’s inequality, and (112) and (113) follow from (98) and (110), respectively. Therefore, there exists some \( N_{e,2} \in \mathbb{N} \), such that, for any \( N > N_{e,2} \),

\[
\rho_{2[N]}(N) \leq e^{-N(\rho_e)}. \tag{115}
\]

When \( \rho_{2[N]}(N) < 1 \), \([Q_{2[N]}^N N > 1, \mathbb{P}[Q_{2[N]}^N \geq 1] = \rho_{2[N]}(N) \). So, for any \( N > N_{e,2} \), (104) holds.

Next, we prove (105). Consider those \( N > N_e := \max\{N_{e,1}, N_{e,2}\} \). We construct an auxiliary sequence as an upper bound for the original sequence of systems in a similar way as we do in the proof of Theorem 1. Specifically, for any \( N > N_e \),

- for \( i = 1, \ldots, [N^\eta] \), let all type \( i \) jobs have size \( \xi_i(N) := s_i(N)\rho_i (\lambda^N p_i)^{-1} \), and
- for \( i = [N^\eta] + 1, \ldots, 2[N^\eta] - 1 \), let all type \( i \) jobs have size \( \xi_i(N) := s_i(N)\rho_i (\lambda^N p_i)^{-1} \),

and everything else remains the same as the original sequence.

As a consequence, the \( 2[N^\eta] - 1 \) independent subsystems in each auxiliary system are all M/D/\( \cdot \) queues and they have a common traffic intensity \( \rho_e \). Next, we show that

\[
\text{for } i = 1, \ldots, [N^\eta], \quad \xi_i(N) \geq f_i(N) = i, \tag{116}
\]

\[
\text{for } i = [N^\eta] + 1, \ldots, 2[N^\eta] - 2, \quad \xi_i(N) \geq f_i(N), \tag{117}
\]

\[
\xi_{2[N^\eta] - 1}(N) \geq f_{2[N^\eta] - 1}(N). \tag{118}
\]

To prove (116) for \( i = 1, \ldots, [N^\eta] \), we use (19) to calculate the traffic intensity of the \( i \)th subsystem in the \( N \)th original system:

\[
\rho_i(N) = \frac{\lambda^N p_i i}{s_i(N)} = \rho_i e^{-Np_i \mu(1 - \sigma(e))}, \tag{119}
\]

and therefore

\[
i = s_i(N)\rho_i (\lambda^N p_i)^{-1} \cdot \frac{Np_i \mu(1 - \sigma(e))}{s_i(N)\rho_i (\lambda^N p_i)^{-1}} \leq s_i(N)\rho_i (\lambda^N p_i)^{-1}, \tag{120}
\]

which gives (116). We then turn to proving (117). For \( i = [N^\eta] + 1, \ldots, 2[N^\eta] - 2 \), the traffic intensity of the \( i \)th subsystem in the \( N \)th original system satisfies

\[
\rho_i(N) \leq \frac{\lambda^N p_i f_i(N)}{s_i(N)} = \rho_i e^{-Np_i \mu f_i(N) \mu(1 - \sigma(e))} \leq \rho_e, \tag{121}
\]

where the first inequality holds because all type \( i \) jobs’ sizes in the \( N \)th original system are no greater than \( f_i(N) \), and the equality in (121) follows from applying (17). Then, (117) immediately follows from (121). To establish (118), we bound the traffic intensity of the \( (2[N^\eta] - 1) \)-th subsystem in the \( N \)th original system:

\[
\rho_{2[N^\eta] - 1}(N) \leq \frac{\lambda^N p_i f_{2[N^\eta] - 1}(N)}{s_i(N)} \tag{122}
\]

\[
= \frac{\lambda^N p_i f_{2[N^\eta] - 1}(N)}{[Np_i f_{2[N^\eta] - 1}(N) \mu(1 - \sigma(e))]} \tag{123}
\]

\[
= \rho_i e^{-Np_i f_{2[N^\eta] - 1}(N) \mu(1 - \sigma(e))} \tag{124}
\]

where (122) is because of the fact that all type \( 2[N^\eta] - 1 \) jobs’ sizes in the \( N \)th original system are no greater than \( f_{2[N^\eta] - 1}(N) \), and (123) follows from (92). From (122) \( \leq (124) \), we obtain that

\[
f_{2[N^\eta] - 1}(N) \leq s_{2[N^\eta] - 1}(N)\rho_e (\lambda^N p_i f_{2[N^\eta] - 1}(N))^{-1} \cdot \frac{Np_i f_{2[N^\eta] - 1}(N) \mu(1 - \sigma(e))}{[Np_i f_{2[N^\eta] - 1}(N) \mu(1 - \sigma(e))]} \tag{125}
\]

which implies (118).

By the construction of the auxiliary sequence (116)-(118), the only difference between a subsystem in the \( N \)th system of the auxiliary sequence and the corresponding subsystem in the \( N \)th system of the original sequence is that each job in the former can have a greater size (but never a smaller size). Therefore

\[
Q_{i,e}^N \leq \alpha Q_{i,e,u}^N \quad \text{for all } i = 1, \ldots, 2[N^\eta] - 1, \tag{126}
\]
where \( Q_{i,\epsilon,u}^{N} \) denotes the steady-state total number of type \( i \) jobs in the \( N \)th system of the auxiliary sequence, and hence
\[
\sum_{i=1}^{2[N^q]−1} Q_{i,\epsilon,u}^{N} \leq u \sum_{i=1}^{2[N^q]−1} Q_{i,\epsilon,u}^{N}.
\] (127)

Define
\[
\delta_N := (s_1(N), \ldots, s_{2[N^q]−1}(N)).
\] (128)

Then, for any \( N > N_\epsilon \), it follows from (127) that
\[
P\left\{ \sum_{i=1}^{2[N^q]−1} Q_{i,\epsilon,u}^{N} \geq N − 1 \right\}
\leq P\left\{ \sum_{i=1}^{2[N^q]−1} Q_{i,\epsilon,u}^{N} \geq N − 1 \right\}
= \sum_{j=0}^\infty P\left\{ \sum_{i=1}^{2[N^q]−1} Q_{i,\epsilon,u}^{N} = N − 1 + j \right\}
\leq \sum_{j=0}^\infty \left( (N − 1 + j) + (2[N^q] − 1) - 1 \right) K(\rho_e)^{2[N^q]−1} \rho_e^j
\times P\{Q_{i,\epsilon,u}^{N} \geq s_i(N)\}, \text{ for all } i = 1, \ldots, 2[N^q] − 1
\leq \sum_{j=0}^\infty \left( (N + 2[N^q] + j - 3) \right) \frac{K(\rho_e)^{2[N^q]−1}}{2[N^q] - 1} \prod_{i=1}^{2[N^q]−1} P\{Q_{i,\epsilon,u}^{N} \geq s_i(N)\}
\leq \frac{(eN)^{2[N^q]−2}}{1 - \rho_e} \prod_{i=1}^{2[N^q]−1} P\{Q_{i,\epsilon,u}^{N} \geq s_i(N)\}[1 + o(1)],
\] (129)

where (129) follows from (70) and Lemma 3, (130) is due to (71) and the independence among \( Q_{i,\epsilon,u}^{N} \)'s, and (131) holds by Lemma 4. Now, we apply Lemma 2 to (131) and obtain the following upper bound:
\[
P\left\{ \sum_{i=1}^{2[N^q]−1} Q_{i,\epsilon,u}^{N} \geq N − 1 \right\}
\leq \frac{(eN)^{2[N^q]−2}}{1 - \rho_e} \prod_{i=1}^{2[N^q]−1} P\{Q_{i,\epsilon,u}^{N} \geq s_i(N)\}[1 + o(1)],
\] (132)

which yields (105), upon taking the logarithm, dividing both sides by \( N \), and letting \( N \to \infty \). Finally, combining (104) and (105) yields (106), which together with (12) implies (2). This establishes the weak optimality. \( \square \)

6. General job sizes. In this section, we prove Theorem 3, which is on systems with general job size distributions.

Proof of Theorem 3. By the continuity of \( I(\cdot) \), (20) and the fact that \( E[S_{\delta_0}] \in [E[S], E[S] + \delta_0] \) for any \( \epsilon > 0 \), there exists \( \delta_0 = \delta(\epsilon) > 0 \) such that (21) holds. Consider a sequence of queues indexed by the number of servers; namely, \( \delta_0 \)-systems, with the same parameters as the original sequence in the QD regime, except that the job sizes are equal in distribution to \( S_\delta \) as defined in (20). Because \( S_{\delta_0} \) is a random variable whose possible values have the same divisor \( \delta_0 \) and \( E[S_{\delta_0}] \leq E[S + \delta_0]^a < \infty \), then for \( \epsilon > 0 \), we can find a policy \( \pi_{\epsilon,S}(N) := \pi_{\delta_0}(N, S_{\delta_0}) \) as given by Corollary 1 such that
\[
\limsup_{N \to \infty} \frac{1}{N} \log P\{Q_{\epsilon,S,\delta_0}^{N} \geq N\} < -I(\rho_{\delta_0}) + \frac{1}{2} \epsilon,
\] (133)

where \( Q_{\epsilon,S,\delta_0}^{N} \) denotes the steady-state total number of jobs in the \( N \)th \( \delta_0 \)-system under policy \( \pi_{\epsilon,S}(N) \).

For any \( N \), consider the original system with job size \( S \) and the \( \delta_0 \)-system with job size \( S_{\delta_0} \), both under policy \( \pi_{\epsilon,S}(N) \). The routing of jobs and the server allocation are the same in these two systems, and their only difference is that some jobs in the \( \delta_0 \)-system have a greater size than the corresponding jobs in the original system. Let \( Q_{\epsilon}^{N} \) denote the steady-state total number of jobs in the original \( N \)-server system under policy \( \pi_{\epsilon,S}(N) \). From the foregoing argument, it follows that \( Q_{\epsilon}^{N} \leq u Q_{\epsilon,S,\delta_0}^{N} \), and therefore
\[
P\{Q_{\epsilon}^{N} \geq N\} \leq P\{Q_{\epsilon,S,\delta_0}^{N} \geq N\}.
\] (134)
This, together with (133) yields
\[
\limsup_{N \to \infty} \left( \frac{1}{N} \log P \{ Q_e^N \geq N \} \right) < - I(\rho_{\theta_0}) + \frac{1}{2} \varepsilon. \tag{135}
\]
Finally, combining (135) with (21), we obtain (2) and establish the weak optimality. □

7. Concluding remarks. This paper focuses on many-server asymptotic analysis of a scheduling policy widely used and studied for computer systems; namely, SITA. Our main result is the construction of SITA policies that are capable of mitigating the impact of job size variability on system performance, as formally characterized by our weak optimality notion. This result provides positive theoretical support for using SITA, regardless of the job size variability level.

The main technical merit of our result lies in the generality of the job size distribution in our model, more specifically, our extension from finite support to unbounded support job sizes. This extension is mathematically challenging and there is no existing methodology for such extensions in the queueing theory or computer performance analysis literature. The present paper therefore contributes to the literature by illustrating one approach to task assignment in the presence of an unbounded job size support and also the corresponding steady-state performance analysis.

Also, our main analytical result is the first on the steady-state performance of the M/G/N queue under any scheduling discipline in the QD regime. The key technical challenge that we address in the performance analysis of the proposed SITA policy \( \pi_e(N) \) in an \( N \)-server system is the identification of an auxiliary queuing system that meets three criteria: the auxiliary system (1) serves as a performance upper bound with respect to the chosen metric for the original \( N \)-server system under \( \pi_e(N) \), (2) is easier to analyze than the original system, and (3) performs sufficiently close to the lower bound system (i.e., the infinite-server queue). Specifically, our upper bound system consists of \( m(N) \) load-balanced M/D/\( \eta \) subsystems, where \( m(N) = 2^{\lceil N^\eta \rceil} \) for some \( \eta \in (0, 1) \), and a single-server queue. To analyze \( m(N) \) M/D/\( \eta \) queues in parallel, we have developed several new results on the M/D/\( c \) queue (Theorem 5, Lemmas 1–3) and used existing combinatorial results (i.e., (70) and (71)). In addition, this upper bound system performs \( \varepsilon \)-close—in the sense of Definition 2—to the infinite-server queue due to our careful construction of the policy, such as the parameterization by the moment index of the job size distribution.

Our main result, however, only states that if the proposed SITA policy is adopted, then weak optimality is achieved. We are not able to prove or disprove that such job separation is necessary. In the literature there are asymptotic results on the queue length process in the G/G/N queue, but the steady-state analysis for G/G/N or even M/G/N queues is still an open problem. We conjecture that FCFS achieves strong optimality and may even achieve exact asymptotic coincidence in the sense of (28), for a fairly general class of job size distributions.

On the other hand, the variability level of the job size distribution can influence the performance of a scheduling policy, especially one that does not exploit the information of the job size distribution such as FCFS. In fact, the disparity in performance between heavy-tailed and light-tailed cases has been observed in single-server queues (see discussion in Wierman and Zwart [33, §1] for more details). Therefore the above conjecture may only hold true under certain job size distribution assumptions, which can be weaker or stronger than the assumption we have made in this study, i.e., finite \( \alpha \)th moment for some \( \alpha > 1 \).

Acknowledgments. This research is partly supported by National Science Foundation [Grants 0727400, 0805979, and 1030589], an IBM faculty award, and a VIDI grant from The Netherlands Organisation for Scientific Research.

Appendix A. Proof of Lemma 3. For any \( \bar{r}, \bar{s} \in \mathbb{Z}_+^m \), we define
\[
J_1(\bar{r}, \bar{s}) := \{ i \in [1, \ldots, m]: r_i < s_i \},
\]
\[
J_2(\bar{r}, \bar{s}) := \{ i \in [1, \ldots, m]: r_i > s_i \},
\]
\[
J_3(\bar{r}, \bar{s}) := \{ i \in [1, \ldots, m]: r_i = s_i \},
\]
\[
d_k(\bar{r}, \bar{s}) := \sum_{i \in J_k(\bar{r}, \bar{s})} |r_i - s_i| \quad \text{for } k = 1, 2.
\]
For any \( \bar{n} \in S(\bar{c}, j) \), we have

\[
\mathbb{P}\{Q_i = n_i, i = 1, \ldots, m\} = \mathbb{P}\{Q_i = n_i, i \in J_1(\bar{n}, \bar{c})\} \times \mathbb{P}\{Q_i = n_i, i \in J_2(\bar{n}, \bar{c})\} \leq \rho_n^{d_1(\bar{n}, \bar{c})} \times \mathbb{P}\{Q_i = n_i, i \in J_1(\bar{n}, \bar{c})\} \times \mathbb{P}\{Q_i = n_i, i \in J_2(\bar{n}, \bar{c})\}
\]

(A.136)

\[
\leq \sum_{i=1}^n \mathbb{P}\{Q_i = c_i, i \in J_1(\bar{n}, \bar{c})\} \times K(\rho_n)^{d_2(\bar{n}, \bar{c})} \times \mathbb{P}\{Q_i = c_i, i \in J_2(\bar{n}, \bar{c})\} \leq K(\rho_n)^m \mathbb{P}\{Q_i = c_i, \text{ for all } i = 1, \ldots, m\},
\]

(A.137)

where (A.136) holds by independence of the \( m \) queues and (A.137) follows from Lemma 1. Note that, in (A.137), \( J_1(\bar{n}, \bar{c}) \) denotes the cardinality of \( J_1(\bar{n}, \bar{c}) \), which is at most \( m \), and \( d_2(\bar{n}, \bar{c}) - d_1(\bar{n}, \bar{c}) = j \) for any \( \bar{n} \in S(\bar{c}, j) \).

Appendix B. Proof of Lemma 4

\[
\sum_{j=0}^{\infty} \frac{(N+2(N^2)+j-3)e}{2(N^2)-2} e^{2N^2-2} \leq \sum_{j=0}^{\infty} e^{2N^2-2} \left( 1 + \frac{N+j-1}{2(N^2)-2} \right)^{2N^2-2} \rho_e^j \leq \sum_{j=0}^{\infty} e^{2N^2-2}(N+j)^{2N^2-2} \rho_e^j
\]

for all \( N \in \mathbb{N} \), such that \( 2(N^2) \geq 1 \). We only consider such \( N \)'s and show that

\[
\sum_{j=0}^{\infty} e^{2N^2-2}(N+j)^{2N^2-2} \rho_e^j = \frac{(eN)^{2N^2-2}}{1-\rho_e}[1+o(1)].
\]

(B.138)

For convenience, we define the ratio between the left-hand side and the right-hand side of (B.138):

\[
R := \frac{\sum_{j=0}^{\infty} e^{2N^2-2}(N+j)^{2N^2-2} \rho_e^j}{(eN)^{2N^2-2}/(1-\rho_e)} = \sum_{j=0}^{\infty} \left( 1 + \frac{j}{N} \right)^{2N^2-2} \rho_e^j(1-\rho_e),
\]

(B.139)

which can be bounded from below and from above as follows:

\[
1 = \sum_{j=0}^{\infty} \rho_e^j(1-\rho_e) \leq R \leq \sum_{j=0}^{\infty} e^{j/N} (2N^2-2) \rho_e^j(1-\rho_e) = \sum_{j=0}^{\infty} (e^{2N^2-2}j/N \, \rho_e^j(1-\rho_e).
\]

(B.140)

We take \( N \) large enough such that \( e^{2N^2-2}j/N \rho_e < 1 \) and then have

\[
1 \leq R \leq \frac{1-\rho_e}{1-e^{2(N^2)-2}j/N \rho_e}.
\]

Letting \( N \to \infty \) yields the desired limit.

References