

Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime: Asymptotics of the stationary distribution

David Gamarnik
Operations Research Center
and Sloan School of Management
100 Main Street
MIT, Cambridge, MA, 02139
gamarnik@mit.edu *

Alexander L. Stolyar
Bell Labs, Alcatel-Lucent
600 Mountain Ave., 2C-322
Murray Hill, NJ 07974
stolyar@research.bell-labs.com †

February 28, 2012

Abstract

We consider a heterogeneous queueing system consisting of one large pool of $O(r)$ identical servers, where $r \rightarrow \infty$ is the scaling parameter. The arriving customers belong to one of several classes which determines the service times in the distributional sense. The system is heavily loaded in the Halfin-Whitt sense, namely the nominal utilization is $1 - a/\sqrt{r}$ where $a > 0$ is the spare capacity parameter. Our goal is to obtain bounds on the steady state performance metrics such as the number of customers waiting in the queue $Q^r(\infty)$. While there is a rich literature on deriving process level (transient) scaling limits for such systems, the results for steady state are primarily limited to the single class case.

This paper is the first one to address the case of heterogeneity in the steady state regime. Moreover, our results hold for any service policy which does not admit server idling when there are customers waiting in the queue. We assume that the interarrival and service times have exponential distribution, and that customers of each class may abandon while waiting in the queue at a certain rate (which may be zero). We obtain upper bounds of the form $O(\sqrt{r})$ on both $Q^r(\infty)$ and the number of idle servers. The bounds are uniform w.r.t. parameter r and the service policy. In particular, we show that $\limsup_r \mathbb{E} \exp(\theta r^{-\frac{1}{2}} Q^r(\infty)) < \infty$. Therefore, the sequence $r^{-\frac{1}{2}} Q^r(\infty)$ is tight and has a uniform exponential tail bound. We further consider the system with strictly positive abandonment rates, and show that in this case every weak limit $\hat{Q}(\infty)$ of $r^{-\frac{1}{2}} Q^r(\infty)$ has a sub-Gaussian tail. Namely $\mathbb{E}[\exp(\theta(\hat{Q}(\infty))^2)] < \infty$, for some $\theta > 0$.

1 Introduction

Recently we have witnessed an explosion of interest in large scale queueing systems operating in the so-called Halfin-Whitt heavy traffic regime. Such systems are attractive for modeling large scale service systems such as call/contact services or computer farms (used in cloud computing), as they achieve high utilization and low service delay at the same time [AAM07], [BGM⁺05], [NGM03]. Most of the prior literature on the subject, including the paper by Halfin and Whitt [HW81], which has initiated the subject, focused on the single class case where customers with identical in distribution service requirements are processed by a large pool of identical servers. This includes the earlier follow up papers such as [RP00], which

*Research supported in part by the NSF grants CMMI-0726733.

†The authors wish to thank the Newton Institute of Mathematics, Cambridge, UK, where this work was initiated and partly conducted, for their hospitality.

focused on the case of exponential and phase-type distribution of model primitives, as well as more recent works [Ree09], [JMM04], [MM05], [GM08], [KR], [KR01], [GG11].

The multiclass models have been analyzed as well [DT10], [ST10], [AG], [AMR04a], [AMR04b], specifically in the context of finding asymptotically optimal control policies, but primarily in the transient as opposed to the steady state regime. In the multiclass environment it is assumed that the processing times may depend on the customer class. (Furthermore, in more general models with heterogeneous servers, the processing times may depend on customer class and server type combination.) The lack of literature analyzing multiclass models in steady state is unfortunate, since multiclass systems offer a richer set of models capturing, for example, call centers with agent skill and customer requirement heterogeneity. The results obtained for single class models for the steady state regimes, unfortunately, are not applicable in the multiclass setting. The original paper by Halfin and Whitt [HW81] relies on the Erlang's formula for the $M/M/N$ queueing systems. [RP00] only analyze the transient regime. The remaining aforementioned papers [Ree09], [JMM04], [MM05], [GM08], [KR], [KR01], [GG11] covering the steady state regime also use the single class assumption in an important way. Thus a new set of tools need to be developed to address heterogeneous models.

In this paper we consider a multiclass queueing system with $O(r)$ parallel identical servers in the Halfin-Whitt regime, where parameter $r \rightarrow \infty$ and the system (nominal) utilization scales as $1 - O(1/\sqrt{r})$. The multiplicity of classes comes from heterogeneity of customer mean service times. Namely, customers class i have the mean service requirement $1/\mu_i$. The arrival processes of different types are assumed to be independent Poisson (of rate $O(r)$), and the service time distributions are assumed to be independent exponential. Class i customers waiting in the queue abandon the system at rate $\nu_i \geq 0$, also in the Markov fashion. An arbitrary non-idling, state dependent (in a sense to be explained) customer service (scheduling) policy is implemented. Interestingly, the details of this policy are irrelevant for our results. Moreover, it is allowed to have a different policy for different values of parameter r . Our results establish, in particular, that the steady state (random) queue length denoted by $Q^r(\infty)$ scales like $O(\sqrt{r})$ as $r \rightarrow \infty$. To be precise, we show that for some fixed parameter $\theta > 0$, $\limsup_{r \rightarrow \infty} \mathbb{E} \left[\exp \left(\theta r^{-\frac{1}{2}} Q^r(\infty) \right) \right] < \infty$, which in particular ensures that the sequence of distributions of $(1/\sqrt{r})Q^r(\infty)$ is tight. If in addition, the scheduling policy admits a process level limit $\hat{Q}(t)$ of the processes $\hat{Q}^r(t) = (1/\sqrt{r})Q^r(t)$, then our results imply that any weak limit (along a subsequence) of the sequence of stationary distributions of \hat{Q}^r is the stationary distribution of \hat{Q} ; in other words, we establish an interchange of heavy traffic and steady state limits. Thus our paper continues an earlier stream of works establishing interchange of limits for queueing networks [GZ06], [BL09], [GM08], [FPG08], [Kat10].

Additionally, for the case when all abandonment rates ν_i are strictly positive, we provide a sub-Gaussian bound on the tail of $\hat{Q}(\infty)$, where $\hat{Q}(\infty)$ is any weak limit of $\hat{Q}^r(\infty) = (1/\sqrt{r})Q^r(\infty)$. A lot of recent literature has been focused on queueing models with abandonments [GMR02], [DH10], [MM]. It was observed that at a process level the presence of abandonments leads to a Ornstein-Uhlenbeck type behavior for the limiting process. Thus one would expect that the steady state measures (of the limit process) should have a sub-Gaussian tail. This was explicitly conjectured in [DH10]. In this paper we qualitatively confirm the conjecture. One might naively conjecture, that in parallel to our exponential-tail result described earlier, one would have $\limsup_r \mathbb{E} \left[\exp \left(\theta (\hat{Q}^r(\infty))^2 \right) \right] < \infty$, for some small enough θ . But this does not hold even in the single-class model, and we provide a simple illustration why, later in the paper. However, we do show that any weak limit $\hat{Q}(\infty)$ of the sequence $\hat{Q}^r(\infty)$ satisfies $\mathbb{E}[\exp(\theta(\hat{Q}(\infty))^2)] < \infty$ for small enough $\theta > 0$. Namely the weak limits of the (scaled) queue length indeed have sub-Gaussian tails.

The remainder of the paper is organized as follows. The model is described and the main results are stated in the next section. In Section 3 we obtain some preliminary results. Specifically, we obtain some useful bounds that follow from the lower bounds derived by comparing with the $M/M/\infty$ model, and also a monotonicity (with respect to abandonment rates) result. Our first main result on the uniform exponential-tail bounds is proved in Sections 4 and 5. Another main result – the sub-Gaussian-tail bounds on the weak limits of stationary distributions, in the case of non-zero abandonment rates – is proved in Section 6.

We close this section with some notational conventions. For any real number b , let $b^+ = \max(b, 0)$, $b^- = -\min(b, 0)$, $\lfloor b \rfloor$ be the largest integer not greater than b , and $\lceil b \rceil$ be the smallest integer not less than b .

$\mathbb{R}, \mathbb{R}_+, \mathbb{Z}, \mathbb{Z}_+$ denote the set of real values, non-negative real values, integer values and non-negative integer values, respectively. \Rightarrow denotes convergence in distribution. $I\{\cdot\}$ is the indicator function taking value 1 when the event inside $\{\cdot\}$ takes place, and 0 otherwise.

2 Model and results

We consider a sequence of queueing systems, indexed by parameter $r = 1, 2, \dots$, increasing to infinity. The system with index r has $N^r = r + a\sqrt{r}$ identical servers, for some fixed $a > 0$. (To be precise, this quantity needs to be integer, so it should be, for example, $\lceil r + a\sqrt{r} \rceil$. This subtlety does not create any difficulties, besides clogging notation. To improve exposition, let us assume that $r + a\sqrt{r}$ itself is integer.) We assume that there is a finite set I of Poisson input customer flows i called classes with rates $\lambda_i r, i \in I$. Each customer of class $i \in I$ has an exponentially distributed service time (requirement) with mean $1/\mu_i$, where (the service rate) parameter $\mu_i > 0$. At any given time a customer is either in service (by one server) or is waiting in the queue of unlimited capacity; each server can serve at most one customer at any given time. If and when the cumulative amount of time the customer spends in service (which can, in principle, be preempted and resumed multiple times) reaches its service requirement, the customer leaves the system. Upon arriving into the system, for each class i customer a random patience time is generated according to the exponential distribution with mean $1/\nu_i$, where (the abandonment rate) parameter $\nu_i \geq 0$. If $\nu_i = 0$, the patience time is infinite. If and when the cumulative amount of time the customer spends in the queue (a customer can, in principle, move between queue and service multiple times) reaches its patience time, the customer abandons and leaves the system. Speaking less formally, a class i customer leaves the system in a small time interval dt with probability $\mu_i dt$ if it is in service, and with probability $\nu_i dt$ if it is in the queue.

We assume that $\sum_i \lambda_i / \mu_i = 1$. Then the nominal utilization of the system is

$$\frac{\sum_i \lambda_i r / \mu_i}{r + a\sqrt{r}} = 1 - \frac{a}{\sqrt{r} + a} = 1 - \frac{a}{\sqrt{r}} - o\left(\frac{1}{\sqrt{r}}\right),$$

namely the queueing system is in the Halfin-Whitt heavy traffic regime. Notice that when all abandonment rates $\nu_i = 0$, the nominal utilization is the actual average utilization. Introduce the following notations: $\rho_i \triangleq \lambda_i / \mu_i$ (implying $\sum_i \rho_i = 1$), $\mu_{\min} \triangleq \min_i \mu_i$ and $\mu_{\max} \triangleq \max_i \mu_i$, $\nu_{\min} \triangleq \min_i \nu_i$ and $\nu_{\max} \triangleq \max_i \nu_i$. For every time $t \in \mathbb{R}_+$, let $Z_i^r(t)$ be the number of type i customers in the system both in service and waiting in the queue, and $\Psi_i^r(t)$ be the number of type i customers in service at time t . Then $Q_i^r(t) = Z_i^r(t) - \Psi_i^r(t)$ is the number of type i customers waiting in the queue at time t .

We now formally define a family of service disciplines under the consideration. Rather than specifying this family directly we describe the properties we want this family to satisfy. Namely, we take a general point of view that the system evolution is described by some countable continuous time Markov chain, which has a projection $((Z_i^r(t), i \in I), (\Psi_i^r(t), i \in I))$, and which satisfies Assumption 1 below. We then demonstrate that under many interesting policies, the corresponding Markov processes indeed satisfy this assumption.

All formal results and proofs in this paper are concerned with abstract processes satisfying Assumption 1. Thus, the physical notions of customer arrivals and departures, introduced earlier in this section, can be viewed as no more than “interpretations” of increments and decrements of components Z_i of the process. But, of course, it will be useful to keep such physical interpretations in mind, as they provide important insights into our methods.

Assumption 1. *Assume that for each r there exists a Markov process $(X^r(t), t \in \mathbb{R}_+)$ in a countable state space \mathcal{X}^r with the following properties:*

1. Vector $((Z_i^r(t), i \in I), (\Psi_i^r(t), i \in I)) \in \mathbb{Z}_+^{2|I|}$, such that $\Psi_i^r(t) \leq Z_i^r(t)$ for all i , is a deterministic function f of $X^r(t)$; this function is such that for every value \mathbf{w} within its range, the pre-image $f^{-1}(\mathbf{w})$ is finite cardinality. Vector $(Q_i^r(t), i \in I) \in \mathbb{Z}_+^{|I|}$ is defined by $Q_i^r(t) = Z_i^r(t) - \Psi_i^r(t)$, and

therefore is a deterministic function of $X^r(t)$ as well.

Interpretation: The process $(Z_i^r(t), \Psi_i^r(t), i \in I)$, which describes the performance of the system we are interested in, is a projection of some “more detailed” underlying Markov process $X^r(t)$. The number of states of $X^r(t)$ that correspond to any given value of $(Z_i^r(t), \Psi_i^r(t), i \in I)$ is finite.

2. The following relation holds:

$$\sum_{i \in I} \Psi_i^r(t) = \min\{N^r, \sum_{i \in I} Z_i^r(t)\}. \quad (1)$$

Interpretation: Whatever scheduling policy the process $X^r(t)$ “encodes”, it is non-idling.

3. The transition rates of the process $X^r(t)$ are such that, for any state $X^r(t)$:

- (a) The total rate of all transitions such that vector $(Z_i^r(t), i \in I)$ changes to $(Z_i^r(t), i \in I) + \mathbf{e}_\ell$ is $\lambda_\ell r$, where \mathbf{e}_i is the i -th unit vector in $\mathbb{R}^{|I|}$.
- (b) If $Z_\ell^r(t) \geq 1$, then the total rate of all transitions such that vector $(Z_i^r(t), i \in I)$ changes to $(Z_i^r(t), i \in I) - \mathbf{e}_\ell$ is $\mu_\ell \Psi_\ell^r(t) + \nu_\ell Q_\ell^r(t)$.
- (c) The transitions such that any component of $(Z_i^r(t), i \in I)$ increases or decreases by more than 1, as well as transitions changing more than one component of $(Z_i^r(t), i \in I)$, have zero rates.

Interpretation: Scheduling policy is consistent with the notions of input processes, service times and patience described at the beginning of this section. Note that there might be positive rate transitions that “reshuffle” the values of occupancies $(\Psi_i^r(t), i \in I)$ in an arbitrary way, as long as conditions (1) and $\Psi_i^r(t) \leq Z_i^r(t)$ are not violated.

4. Markov process $X^r(t)$ is irreducible.

Interpretation: This condition is for convenience only – it allows us to claim not only existence, but also uniqueness of stationary distribution (as discussed below). Also, this condition is very non-restrictive – typically the “encoding process” $X^r(t)$ can be constructed in a way such that there is only one state corresponding to the “empty” system (with all $Z_i^r = 0$), and this state obviously is reachable from any other.

Assumption 1 encompass a broad variety of non-idling disciplines, with or without service preemptions. For illustrative purposes, let us discuss the Class Priority (CP) and First-In-First-Out (FIFO) examples. The CP policy is fixed by setting a total order \prec on classes which without the loss of generality is assumed to be $1 \prec 2 \prec \dots \prec |I|$. This means that the jobs in class $|I|$ have the highest priority, the jobs in class $|I| - 1$ have the next highest, etc. At any time, the servers give priority to the jobs with the highest priority in a non-preemptive or preemptive fashion. Namely, every server upon service completion inspects the queue and selects for service a job with the highest priority level, ties broken arbitrarily. For the non-preemptive version, once the service is initiated it is completed without interruption; in this case, we can simply take $X^r(t) = (Z_i^r(t), \Psi_i^r(t), i \in I)$. For the preemptive version, if an arriving customer of type i finds all servers busy and if $j < i$, where j is the lowest priority being served, the arriving customer will force one of the type j customers out to the queue and replace it in service; in this case, the process is even simpler, $X^r(t) = (Z_i^r(t), i \in I)$, because $(\Psi_i^r(t), i \in I)$ is uniquely determined by $(Z_i^r(t), i \in I)$. In either case, the “empty” state, with all $Z_i^r(t) = \Psi_i^r(t) = 0$, is reachable from any other, and therefore the chain is irreducible.

For the FIFO policy, the state space \mathcal{X}^r of chain $X^r(t)$ is the set of finite sequences $x_1 x_2 \dots x_m$, where $x_n \in I$. Each sequence represents the set of customers in the system, in the order of their arrivals. The first $\min\{m, N^r\}$ customers are being served, and the remaining ones wait in queue. Clearly, $(Z_i^r(t), \Psi_i^r(t), i \in I)$ is uniquely determined by $X^r(t)$. The empty state is reachable from any other.

We will now discuss several general properties of the process $X^r(t)$ that follow directly from Assumption 1. For $\mathbf{x} \in \mathcal{X}^r$, denote by $T(\mathbf{x})$ the set of states to which a direct transition from \mathbf{x} is possible, and by $T^{-1}(\mathbf{x})$ the set of states from which a direct transition to \mathbf{x} is possible. It is easy to observe that for any \mathbf{x} both $T(\mathbf{x})$ and $T^{-1}(\mathbf{x})$ are finite cardinality – this follows from the fact that there is only a finite number of states

corresponding to each $((Z_i^r(t), i \in I), (\Psi_i^r(t), i \in I))$, and then to each $(Z_i^r(t), i \in I)$, and direct transitions cannot cause any $Z_i^r(t)$ to change by more than 1. We can also observe that, starting any initial state, with probability 1 the process makes at most a finite number of transitions in any finite time interval; indeed, with probability 1 each $Z_i^r(t)$ remains finite on finite intervals, because the total rate of all transitions causing it to increase (by 1) is constant and equal $\lambda_i r$, and then the number of states visited on finite intervals is finite, from which we easily obtain that the number of transitions is finite as well. This means that the process satisfying Assumption 1 is well defined by its transition rates – with probability 1 there are no “explosions”.

Since the system nominal utilization is strictly less than 1 and condition (1) must hold, the total workload (expected unfinished work) $\sum_i Z_i^r(t)/\mu_i$ must have a negative average drift, as long as $\sum_i Z_i^r(t)/\mu_i \geq N^r/\mu_{min}$. (The latter condition implies $\sum_i Z_i^r(t) \geq N^r$, in which case the average rate at which unfinished work in the system is processed is strictly greater than the average rate at which new work arrives.) This, along with the fact that for any $K \geq 0$ there is only a finite number of states with $\sum_i Z_i^r(t) \leq K$, and the assumption that Markov chain $X^r(t)$ is irreducible, implies (using standard methods) that $X^r(t)$ is *positive recurrent with a unique invariant probability distribution* π^r . From now on we write $X^r(\infty)$ for the steady state version of $X^r(t)$. The same applies to the variables determined by $X^r(t)$. For example $Q^r(\infty)$ denotes the steady state queue size.

We remark that the irreducibility assumption is made to simplify the exposition – it is non-essential, in the following sense. If we do not require irreducibility, it is easy to show (again, using standard methods) that Markov chain $X^r(t)$ has at most a finite number of classes of communicating states, reachable w.p.1 from any transient state. Restricted to any of those state classes, $X^r(t)$ is positive recurrent, with unique stationary distribution. Thus, stationary distributions of $X^r(t)$ exist and are characterized as convex combinations of a fixed finite set of distributions. All results of this paper would still hold if instead of *the* stationary distribution we would consider *a* stationary distribution.

Next we discuss some properties related to the (infinitesimal) generator A^r of the Markov process $X^r(t)$. The domain of generator A^r might not contain all functions $F : \mathcal{X}^r \rightarrow \mathbb{R}$. However, the following operator \bar{A}^r is defined for *any* function F :

$$\bar{A}^r F(\mathbf{x}) = \sum_{\mathbf{y} \in T(\mathbf{x})} \xi(\mathbf{x}, \mathbf{y})(F(\mathbf{y}) - F(\mathbf{x})), \quad \mathbf{x} \in \mathcal{X}^r,$$

where $\xi(\mathbf{x}, \mathbf{y})$ is the transition rate from \mathbf{x} to \mathbf{y} . Given properties of our process $X^r(t)$, the following property is easily verified directly: any function F , which is constant everywhere except a finite subset of \mathcal{X}^r , is within the domain of A^r and, moreover,

$$A^r F(\mathbf{x}) = \bar{A}^r F(\mathbf{x}), \quad \forall \mathbf{x},$$

$$\mathbb{E} A^r F(X^r(\infty)) = \mathbb{E} \bar{A}^r F(X^r(\infty)) = 0. \quad (2)$$

Let us additionally denote $\Phi_i^r(t) = Z_i^r(t)/\mu_i$. Namely, $\Phi_i^r(t)$ is the total workload at time t associated with class i jobs - the sum of expected remaining service times of all class i jobs present in the system at time t . The total workload is $\Phi^r(t) = \sum_i \Phi_i^r(t)$.

Now let us introduce the diffusion scaled version of our processes:

$$\hat{Z}_i^r(t) = \frac{Z_i^r(t) - \rho_i r}{\sqrt{r}}, \quad \hat{\Psi}_i^r(t) = \frac{\Psi_i^r(t) - \rho_i r}{\sqrt{r}}, \quad \hat{\Phi}_i^r(t) = \frac{\Phi_i^r(t) - \rho_i r/\mu_i}{\sqrt{r}} = \frac{\hat{Z}_i^r(t)}{\mu_i} \quad \hat{Q}_i^r(t) = \frac{Q_i^r(t)}{\sqrt{r}}.$$

Furthermore, let $\hat{Z}^r(t) = \sum_i \hat{Z}_i^r(t)$, $\hat{\Phi}^r(t) = \sum_i \hat{\Phi}_i^r(t)$, $\hat{Q}^r(t) = \sum_i \hat{Q}_i^r(t)$. Namely, $\hat{Z}^r(t)$ is the diffusion scaled total number of customers in the system, $\hat{\Phi}^r(t)$ is the diffusion scaled total workload in the system, and $\hat{Q}^r(t)$ is the diffusion scaled total queue length. The proofs of our main results (Theorems 2 and 4 below) are based on the Lyapunov function technique. We will be considering Lyapunov functions of the form $\exp(\theta \hat{\Phi}^r)$ and $\exp(\theta (\hat{\Phi}^r)^2)$. The random variables $Z^r(t)$, $\hat{Z}^r(t)$, $\Phi_i^r(t)$, etc., corresponding to the steady

state regime are denoted using the same notation, except ∞ replaces t . For example $\hat{Z}^r(\infty)$ is the steady-state number of customers in the system after the diffusion rescaling. In addition, we write $\hat{Z}^{r,+}(\infty)$ to mean $(\hat{Z}^r(\infty))^+$, $\hat{Z}_i^{r,+}(\infty)$ to mean $(\hat{Z}_i^r(\infty))^+$, etc.

We are now ready to state our main results.

Theorem 2. *The following properties hold.*

(i) *If $\nu_i \leq \mu_i$ for all i (including possibly $\nu_i = 0$ for some i), then for every $\theta > 0$*

$$\limsup_{r \rightarrow \infty} \mathbb{E} \exp \left(\theta \sum_i \hat{Z}_i^{r,-}(\infty) \right) < \infty. \quad (3)$$

(ii) *There exists $\bar{\theta} > 0$ such that for every $\theta \in [0, \bar{\theta}]$*

$$\limsup_{r \rightarrow \infty} \mathbb{E} \exp \left(\theta \sum_i \hat{Z}_i^{r,+}(\infty) \right) < \infty. \quad (4)$$

(iii) *There exists $\bar{\theta}' > 0$ such that for every $\theta \in [0, \bar{\theta}']$*

$$\limsup_{r \rightarrow \infty} \mathbb{E} \exp \left(\theta \sum_i \hat{Z}_i^{r,-}(\infty) \right) < \infty. \quad (5)$$

(iv) *In particular, the sequence $\hat{Z}^r(\infty), r \in \mathbb{Z}_+$ is tight.*

Our proof approach can lead to explicit values of $\bar{\theta}$ and $\bar{\theta}'$, in terms of system parameters. (And we emphasize again that Theorem 2, in particular with those values of $\bar{\theta}$ and $\bar{\theta}'$, applies uniformly to *all* service disciplines satisfying Assumption 1.) While these values can possibly be improved (increased), we do not pursue here a goal of finding their sharp upper bounds.

Clearly, statement (iv) is a corollary from (ii) and (iii), because $|\hat{Z}^r(\infty)| \leq \sum \hat{Z}_i^{r,+}(\infty) + \sum \hat{Z}_i^{r,-}(\infty)$, and then

$$\limsup_{r \rightarrow \infty} \mathbb{E} \exp \left(\theta |\hat{Z}^r(\infty)| \right) < \infty,$$

for sufficiently small $\theta > 0$ by application of the Cauchy-Schwartz inequality. The proof of (i) will follow from a simple comparison to the infinite-server system. The proof of (ii) will rely in essential way on statement (i). Finally, the proof of (iii) will rely on statement (ii).

Theorem 2 shows, in particular, that in the Halfin-Whitt heavy traffic regime the steady state total queue length $Q^r(\infty)$ scales like $O(\sqrt{r})$, because

$$\begin{aligned} \hat{Q}^r(\infty) &= (Z^r(\infty) - r - a\sqrt{r})^+ / \sqrt{r} \\ &\leq (Z^r(\infty) - r)^+ / \sqrt{r} \\ &= \hat{Z}^{r,+}(\infty). \end{aligned}$$

Furthermore, our result implies that $\hat{Q}^r(\infty)$ decays at an exponential rate:

$$\mathbb{P} \left(\hat{Q}^r(\infty) \geq x \right) \leq c_1 \exp(-c_2 x),$$

for some constants $c_1, c_2 > 0$ uniformly in r . The steady state number of idle servers, $(r + a\sqrt{r} - Z^r(\infty))^+$, also scales like $O(\sqrt{r})$, because it is upper bounded by $a\sqrt{r} + \sum_i \hat{Z}_i^{r,-}(\infty)\sqrt{r}$.

The tightness of the sequence $\hat{Z}^r(\infty)$ implies the existence of at least one subsequential weak limit. Namely, there exists a subsequence $r_n, n \in \mathbb{Z}_+$ and a random variable denoted by $\hat{Z}(\infty)$, such that $\hat{Z}^r(\infty) \Rightarrow \hat{Z}(\infty)$. Such a limit is not necessarily unique as it may depend on the details of the scheduling policy. Then, Theorem 2 implies the following

Corollary 3. For a sufficiently small fixed $\theta > 0$ and every weak limit $\hat{Z}(\infty)$ of the sequence $\hat{Z}^r(\infty)$, $\mathbb{E}[\exp(\theta|\hat{Z}(\infty)|)] < \infty$.

Proof. Fix positive $\theta \leq \min(\bar{\theta}, \bar{\theta}')$. Fix arbitrary $k > 0$. If $\hat{Z}^r(\infty)$ converges weakly to $\hat{Z}(\infty)$, then

$$\mathbb{E} \left[\min \left(\exp(\theta|\hat{Z}(\infty)|), k \right) \right] = \lim_{r \rightarrow \infty} \mathbb{E} \left[\min \left(\exp(\theta|\hat{Z}^r(\infty)|), k \right) \right] \leq \limsup_{r \rightarrow \infty} \mathbb{E} \left[\exp(\theta|\hat{Z}^r(\infty)|) \right],$$

where the RHS is finite and does not depend on k . It remains to take the limit of LHS on $k \rightarrow \infty$ and use the Monotone Convergence Theorem. \square

Theorem 2 leaves open the question whether (5) holds for all θ . (Equivalently whether (3) holds also when some $\nu_i > \mu_i$.) We conjecture that it is the case, but we are currently unable to verify this statement.

We now turn to the case when every customer class has strictly positive abandonment rate: $\nu_i > 0$. In this case, even tighter bounds on any weak limit of the stationary distribution can be obtained. It is known that the presence of abandonments changes the properties of the tail of the distribution of $\hat{Z}^r(\infty)$. As an illustration consider a single class model $M/M/N^r$, where the abandonment rate $\nu = \mu$ for simplicity. It is straightforward to see that $Z(\infty)$ has the same distribution as in the $M/M/\infty$ system, namely it is Poisson with mean $\lambda r / \mu = r$. As $r \rightarrow \infty$ the distribution of $\hat{Z}^r(\infty)$ then approaches a standard normal distribution. Thus we might expect the tail $\mathbb{P}(\hat{Z}^r(\infty) > x)$ to decay at the rate $\exp(-O(x^2))$. Namely, one might conjecture the following analogue of Theorem 2 - there exists a sufficiently small $\theta > 0$ such that $\sup_r \mathbb{E}[\exp(\theta(\hat{Z}^r(\infty))^2)] < \infty$. However, this cannot be the case - if random variable G has Poisson distribution with mean $r > 0$, then $\mathbb{E}[\exp(\theta G^2)] = \infty$ for every $\theta > 0$. Nevertheless, we can show that the following analogue of Corollary 3 holds.

Theorem 4. For every subsequential weak limit $\hat{Z}(\infty)$ of $\hat{Z}^r(\infty)$, $r \in \mathbb{Z}_+$, the following properties hold.

(i) If $\nu_i \leq \mu_i$ for all i (including possibly $\nu_i = 0$ for some i), then there exists $\bar{\theta}' > 0$, such that for every $\theta \in [0, \bar{\theta}']$

$$\mathbb{E} \exp \left(\theta \left(\hat{Z}^-(\infty) \right)^2 \right) < \infty. \quad (6)$$

(ii) Suppose $\nu_i > 0$ for all $i \in I$. Then there exists $\bar{\theta} > 0$ such that for every $\theta \in [0, \bar{\theta}]$

$$\mathbb{E} \exp \left(\theta \left(\hat{Z}^+(\infty) \right)^2 \right) < \infty. \quad (7)$$

This result shows that, while the presence of abandonments does not change the tail decay rate of $\hat{Z}^{r,+}(\infty)$ for each r , the decay rate of the tail of the weak limit $\hat{Z}^+(\infty)$ is faster - it is Gaussian rather than simply exponential. This, in particular, implies that

$$\mathbb{P} \left(\hat{Q}(\infty) \geq x \right) \leq c_3 \exp(-c_4 x^2),$$

for some constants $c_3, c_4 > 0$, where $\hat{Q}(\infty)$ is any weak limit of $\hat{Q}^r(\infty)$.

3 Key bounds and monotonicity properties

We begin by establishing a few technical results. For the remainder of the paper we drop for convenience the (∞) and superscript r notations, and simply use $Z, Z_i, \hat{Z}_i, \Phi_i$, etc. to denote our steady-state variables. The same applies to the generator A and operator \bar{A} - the superscript r is dropped.

Lemma 5. Suppose $\nu_i \leq \mu_i$ for all i . Then for any $\theta \geq 0$,

$$\limsup_{r \rightarrow \infty} \mathbb{E} \exp(\theta \hat{Z}_i^-) \leq \exp(\rho_i \theta^2 / 2) + 1, \quad (8)$$

and

$$\limsup_{r \rightarrow \infty} \mathbb{E} \exp(\theta \sum_i \hat{Z}_i^-) \leq \prod_i [\exp(\rho_i \theta^2 / 2) + 1]. \quad (9)$$

Proof. We obtain the required lower bounds by comparing with a system with infinitely many servers. Namely, in addition to the original system, i.e. the process $X(t)$, which defines it, consider the system with the same arrival rates, same mean service times, same (exponential) service time distributions, but with infinite number of servers. Let G_i be the steady-state number of customers in service in this infinite server model. Then, each G_i has Poisson distribution with mean $\rho_i r$, and all G_i 's are independent.

Given the $\nu_i \leq \mu_i$ assumption, clearly we can construct both the original and the infinite-server processes on a common probability space in a way such that, w.p.1 $G_i(t) \leq Z_i(t)$ for all t . That is, we can construct a ‘‘joint’’ Markov process, for which the original process $X(t)$ and the infinite-server process are different projections. Given that $X(t)$ is irreducible and positive recurrent, we can make sure that the constructed joint process is also irreducible and positive recurrent, and thus has a unique stationary distribution. Then, the relation $G_i \leq Z_i$ holds in the stationary regime as well. Then also $\hat{G}_i \leq \hat{Z}_i$, where we denoted $\hat{G}_i = (G_i - \rho_i r) / \sqrt{r}$. Since G_i has Poisson distribution with parameter $\rho_i r$, for any real θ ,

$$\mathbb{E} \exp(\theta \hat{G}_i) = \exp \left[-\theta \rho_i \sqrt{r} - \rho_i r (1 - e^{\theta / \sqrt{r}}) \right] \xrightarrow{r \rightarrow \infty} \exp(\rho_i \theta^2 / 2). \quad (10)$$

Using $\hat{Z}_i \geq \hat{G}_i$, (10) immediately implies, that for any $\theta \geq 0$,

$$\limsup_{r \rightarrow \infty} \mathbb{E} \exp(\theta \hat{Z}_i^-) \leq \limsup_{r \rightarrow \infty} \mathbb{E} \exp(\theta \hat{G}_i^-) \leq \exp(\rho_i \theta^2 / 2) + 1, \quad (11)$$

(where we used $\exp(\theta \hat{G}_i^-) \leq 1 + \exp(-\theta \hat{G}_i)$) and

$$\limsup_{r \rightarrow \infty} \mathbb{E} \exp(\theta \sum_i \hat{Z}_i^-) \leq \limsup_{r \rightarrow \infty} \mathbb{E} \exp(\theta \sum_i \hat{G}_i^-) \leq \prod_i [\exp(\rho_i \theta^2 / 2) + 1]. \quad (12)$$

□

Next we obtain a bound which, roughly speaking, says that, when all $\nu_i \leq \mu_i$, it is unlikely for $\hat{\Phi}$ to be large if \hat{Z} is not large.

Lemma 6. Suppose $\nu_i \leq \mu_i$ for all i . For arbitrary $b \geq 0, \theta_1 > 0$, and $n \in \mathbb{Z}_+$, there exists $C = C(b, \theta_1, n)$ such that for all $\theta \in [0, \theta_1]$,

$$\limsup_{r \rightarrow \infty} \mathbb{E} \left[I\{\hat{Z} \leq b\} \left(\sum_i \hat{Z}_i^- \right)^n \exp(\theta \hat{\Phi}) \right] \leq C. \quad (13)$$

Proof. Inequality

$$(1/\mu_{\min}) \sum_i \hat{Z}_i^+ - (1/\mu_{\max}) \sum_i \hat{Z}_i^- \geq \hat{\Phi} \quad (14)$$

holds always. Condition $\hat{Z} \leq b$ is equivalent to

$$\sum_i \hat{Z}_i^+ - \sum_i \hat{Z}_i^- \leq b. \quad (15)$$

Multiplying (15) by $(1/\mu_{\min})$ and combining with (14), we get

$$\hat{\Phi} \leq (1/\mu_{\min} - 1/\mu_{\max}) \sum_i \hat{Z}_i^- + (1/\mu_{\min})b. \quad (16)$$

Applying bound (16) to the left-hand side of (13), and then using (9), we obtain (13). □

Finally, we will need the following monotonicity result. It says, roughly speaking, that for any well-defined system (i.e., a process satisfying Assumption 1) there always exists another well-defined system with smaller abandonment rates, in which the number of customers of each class is larger. It is important to note that Lemma 7 does *not* imply that this monotonicity works in the “opposite direction”: namely, it does *not* imply that for any system there exists another system with *larger* abandonment rates, in which the number of customers of each class is *smaller*.

Lemma 7. *Consider a process $X(t)$, satisfying Assumption 1 for a given set of parameters $r, a, N, \mu_i, \nu_i, i \in I$. Suppose we have a modified set of parameters $r, a, N, \mu_i, \nu'_i, i \in I$, with $\nu'_i \leq \nu$ for all i . Then there exists a Markov process $X'(t)$ with associated $Z'_i(t), \Psi'_i(t)$, satisfying Assumption 1 with the modified parameter set, and such that the following holds. The processes $X(t)$ and $X'(t)$ can be constructed on a common probability space in a way such that, first, with probability 1 $Z_i(t) \leq Z'_i(t)$ and $\Psi_i(t) \leq \Psi'_i(t)$ for all i and t , and, second, the joint process $(X(t), X'(t))$ is an irreducible positive recurrent Markov chain.*

Note that the lemma claim that the joint process $(X(t), X'(t))$ is an ergodic Markov chain, allows us to conclude that the steady-state versions $X(\infty)$ and $X'(\infty)$ can be constructed on a common probability space in a way such that $Z_i(\infty) \leq Z'_i(\infty)$ and $\Psi_i(\infty) \leq \Psi'_i(\infty)$ for all i .

Before giving a proof, we informally describe the simple intuition behind it. Given a queueing system corresponding to parameters $r, a, N, \mu_i, \nu_i, i \in I$, operating under any scheduling policy so that the Assumption 1 is valid, consider a modified system constructed as follows. Assume for simplicity that $\nu_i > 0$ for each i . Upon every abandonment instance of a system (which happens with rate ν_i for each class i customer waiting in the queue), customers abandon only with probability ν'/ν and with the remaining probability $(1 - \nu'/\nu)$ join a new “virtual” class i pool of customers, which we call class i' , $i \in I$ customers, with some abuse of notations. The decision is done independently for all customers. We assume that the class i' customers for any $i \in I$ have lower priority than all the customers in classes $j \in I$ and priority mechanism is preemptive resume. Thus the portion of the system corresponding to classes $j \in I$ only is the same as the original system. The service times of class i' customers are assumed to be exponentially distributed random variable with parameter μ_i . Additionally, class i' customers abandon with rate ν'_i . It is easy to see that each queued classes i customer abandons with rate ν_i ; and each queued customer of class i or i' abandons with rate ν'_i , *if we do not count an i -to- i' conversion as an abandonment*. This mechanism induces a new queueing system such that $Z_i(t) \leq Z'_i(t) \equiv Z_i(t) + Z_{i'}(t)$ (and similarly for Ψ).

Proof. The states $X'(t)$ of the modified process are defined as augmented states $X(t)$, as follows:

$$X'(t) = [X(t), (Z'_i(t), i \in I)],$$

where $(Z'_i(t), i \in I)$ is a vector such that $Z_i(t) \leq Z'_i(t)$ for all i . (Recall that $X(t)$ uniquely determines the associated $Z_i(t), \Psi_i(t), i \in I$. Therefore, corresponding to any state of $X(t)$ there is a countable set of states of $X'(t)$.) Now, consider an arbitrary fixed function, that maps $X'(t)$ into an integer vector $(\Psi'_i(t), i \in I)$, while satisfying the following conditions: $\Psi_i(t) \leq \Psi'_i(t)$ for all i ; $\sum_i \Psi'_i(t) = \min\{N, \sum_i Z'_i(t)\}$. (The deterministic mapping of $X'(t)$ into $(Z'_i(t), i \in I)$ is the projection.) Recall notation $Q_i(t) = Z_i(t) - \Psi_i(t)$, and denote $Q'_i(t) = Z'_i(t) - \Psi'_i(t)$. Observe that we always have $Q_i(t) \leq Q'_i(t)$; indeed, if $\sum_i \Psi_i(t) < N$ then all $Q_i(t) = 0$, and if $\sum_i \Psi_i(t) = N$, then by construction $\sum_i \Psi'_i(t) \geq \sum_i \Psi_i(t) = N$, which coupled with $\sum_i \Psi'_i(t) \leq N$ implies $\Psi_i(t) = \Psi'_i(t)$ for all i .

We will now define the transition rates of process $X'(t)$; they will be such that the order relations $Z_i(t) \leq Z'_i(t)$ are preserved. We will use time indices $t-$ and t for a state just before and just after a transition at time t . The transitions will be of two types. The first type transitions are “driven” by those of process X , and have the same rates. When X makes a transition from $X(t-)$ to $X(t)$, such that $Z_i(t) = Z_i(t-)$ for all i (no “arrivals” or “departures” in the X -process), or $Z_i(t) = Z_i(t-) + 1$ for one of the i (an “arrival” in the X -process), then $X'(t) = (X(t), (Z'_i(t), i \in I))$, where each $Z'_i(t) = Z'_i(t-) + Z_i(t) - Z_i(t-)$. (Simply put, in X' we make the same change of X -component and increment Z'_i 's in the same way.) When, X makes a transition from $X(t-)$ to $X(t)$, such that $Z_i(t) = Z_i(t-) - 1$ for one of the i (a “departure” in the X -process),

say $i = j$, then $X'(t) = (X(t), (Z'_i(t), i \in I))$, where $Z'_i(t) = Z'_i(t-)$ for all $i \neq j$, and $Z'_i(t)$ is set to $Z'_i(t-)$ or $Z'_i(t-) - 1$ with probabilities

$$p = \frac{\nu_i Q_i(t-)}{\nu_i Q_i(t-) + \mu_i \Psi_i(t-)} \frac{\nu_i - \nu'_i}{\nu_i} \quad \text{and} \quad 1 - p,$$

respectively. (In the expression for p the ν_i in both the numerator and denominator cancels out – we left it there to make the meaning of probability p transparent. In particular, the expression does *not* require that $\nu_i > 0$. We also do not need to “worry” about the case when $\nu_i Q_i(t-) + \mu_i \Psi_i(t-) = 0$, because in this case a “departure of class i ” transition in X is impossible, and we only define p for such transitions.)

The second type transitions are associated with each customer class j (these are the “departures” in the X' -process only); namely, with the rate $\nu'_j(Q'_j(t-) - Q_j(t-)) + \mu_j(\Psi'_j(t-) - \Psi_j(t-))$ the transition from $X'(t-)$ to $X'(t)$ is such that $X(t) = X(t-)$, $Z'_j(t) = Z'_j(t-) - 1$, and $Z'_i(t) = Z'_i(t-)$ for all $i \neq j$.

The Markov chains $X'(t)$ and (its projection) $X(t)$ are constructed on the same probability space. The order relations $Z_i(t) \leq Z'_i(t)$ are preserved for all i at all times, by construction. By construction the process $X(t)$ is defined “as itself”. It is easily checked that the process $X'(t)$ satisfies Assumption 1, including the irreducibility condition. Therefore, $X'(t)$ is positive recurrent. Since $X(t)$ is a projection of $X'(t)$, the process $X'(t)$ itself can be taken as the joint process $(X(t), X'(t))$. The proof is complete \square

4 Proof of Theorem 2(i) and (ii)

Statement (i) follows from Lemma 5.

To prove statement (ii) it suffices by Lemma 7 to consider the case when all $\nu_i = 0$, which is what we assume for the rest of the proof. Consider the process in the stationary regime, i.e. \mathbf{x} is random with the distribution equal to the stationary distribution π of the Markov chain. First, for every state $\mathbf{x} \in \mathcal{X}$ we have

$$\bar{A}\hat{\Phi}(\mathbf{x}) = \frac{1}{\sqrt{r}} \left[\sum_i (\lambda_i r) / \mu_i - \sum_i \Psi_i(\mathbf{x}) \right] = -\hat{Z}_a(\mathbf{x}),$$

where $\hat{Z}_a(\mathbf{x}) = \min\{\hat{Z}(\mathbf{x}), a\}$. We remark that this expression does *not* require chain $X(\cdot)$ to be in steady-state.

Consider a (candidate) Lyapunov function $\exp(\theta\hat{\Phi}(\mathbf{x}))$ for every state $\mathbf{x} \in \mathcal{X}$, where $\theta > 0$ is for a moment arbitrary. We use the Taylor expansion

$$\exp(\theta y) = \exp(\theta x) (1 + \theta(y - x) + (1/2)\theta^2(y - x)^2 \exp(\theta z)),$$

for some z satisfying $|z| \leq |y - x|$. Then $\exp(\theta y) \leq \exp(\theta x) (1 + \theta(y - x) + (1/2)\theta^2(y - x)^2 \exp(\theta|y - x|))$. We will apply this for $x = \hat{\Phi}(\mathbf{x})$, $y = \hat{\Phi}(\mathbf{y})$, where $\mathbf{y} \in T(\mathbf{x})$. We obtain

$$\bar{A} \exp(\theta\hat{\Phi}(\mathbf{x})) \leq \exp(\theta\hat{\Phi}(\mathbf{x})) \left[\theta \bar{A}\hat{\Phi}(\mathbf{x}) + (1/2)\theta^2 \left[\sum_i \lambda_i r + (r + a\sqrt{r})\mu_{\max} \right] (\mu_{\min}^{-1}/\sqrt{r})^2 \exp(\theta\mu_{\min}^{-1}/\sqrt{r}) \right] \quad (17)$$

where we use the fact that $|y - x| \leq 1/(\mu_{\min}\sqrt{r})$, and $\sum_i \lambda_i r + (r + a\sqrt{r})\mu_{\max}$ is a crude upper bound on the maximum transition rate (involving a customer arrival or departure). We obtain

$$\bar{A} \exp(\theta\hat{\Phi}(\mathbf{x})) \leq \exp(\theta\hat{\Phi}(\mathbf{x})) \left[-\theta\hat{Z}_a(\mathbf{x}) + (1/2)\theta^2 c_1 \right], \quad (18)$$

where, uniformly on all $r \geq 1$ and all $\theta \leq 1$,

$$c_1 = \left(\sum_i \lambda_i + (1 + a)\mu_{\max} \right) \mu_{\min}^{-2} \exp(\mu_{\min}^{-1}) \geq \left(\sum_i \lambda_i + (1 + a/\sqrt{r})\mu_{\max} \right) \mu_{\min}^{-2} \exp(\theta\mu_{\min}^{-1}/\sqrt{r}).$$

In the rest of this proof we only consider $\theta \leq 1$.

Let $\hat{\Phi}_{(k)}(\mathbf{x}) \triangleq \min\{\hat{\Phi}(\mathbf{x}), k\}$. Clearly, $\exp(\theta\hat{\Phi}_{(k)}(\mathbf{x}))$ is constant everywhere except a finite subset of state space \mathcal{X} , because (Assumption 1) there is only a finite number of states $\mathbf{x} \in \mathcal{X}$ for which $\sum \hat{Z}_i(\mathbf{x})/\mu_i < k$. Therefore, (2) holds for the function $\exp(\theta\hat{\Phi}_{(k)}(\mathbf{x}))$.

Observe that $\hat{\Phi}(\mathbf{x}) \geq k$ implies $\bar{A} \exp(\theta\hat{\Phi}_{(k)}(\mathbf{x})) \leq 0$, and $\hat{\Phi}(\mathbf{x}) < k$ implies $\bar{A} \exp(\theta\hat{\Phi}_{(k)}(\mathbf{x})) \leq \bar{A} \exp(\theta\hat{\Phi}(\mathbf{x}))$. Then, from (18) we obtain:

$$\bar{A} \exp(\theta\hat{\Phi}_{(k)}(\mathbf{x})) \leq I\{\hat{\Phi}(\mathbf{x}) < k\} I\{\hat{Z}(\mathbf{x}) \geq a\} \exp(\theta\hat{\Phi}(\mathbf{x})) \left[-\theta a + (1/2)\theta^2 c_1 \right] \quad (19)$$

$$+ I\{\hat{Z}(\mathbf{x}) < 0\} \exp(\theta\hat{\Phi}(\mathbf{x})) \left[-\theta\hat{Z}(\mathbf{x}) + (1/2)\theta^2 c_1 \right] \quad (20)$$

$$+ I\{0 \leq \hat{Z}(\mathbf{x}) < a\} \exp(\theta\hat{\Phi}(\mathbf{x})) \left[(1/2)\theta^2 c_1 \right]. \quad (21)$$

For the process in steady state, all terms (19)-(21) have finite expected values for each k and r , including the left-hand side of (19). Indeed, the left- and right-hand sides of (19) are bounded, and (20) and (21) are non-negative and have finite expectations by Lemma 6. We now take the expectation w.r.t. the stationary distribution π on \mathcal{X} of both sides, use $\mathbb{E}[\bar{A} \exp(\theta\hat{\Phi}_{(k)}(\mathbf{x}))] = 0$ (which follows from (2)), and cancel θ . Recalling our notation $\hat{\Phi} = \hat{\Phi}^r(\infty)$, which is the same as $\hat{\Phi}(\mathbf{x})$, where \mathbf{x} is distributed according to the stationary measure π , we obtain

$$\mathbb{E} \left[I\{\hat{\Phi} < k\} I\{\hat{Z} \geq a\} \exp(\theta\hat{\Phi})(a - (1/2)\theta c_1) \right] \leq \quad (22)$$

$$\mathbb{E} \left[I\{\hat{Z} < 0\} \exp(\theta\hat{\Phi}) \left[-\hat{Z} + (1/2)\theta c_1 \right] \right] + \quad (23)$$

$$\mathbb{E} \left[I\{0 \leq \hat{Z} < a\} \exp(\theta\hat{\Phi}) \left[(1/2)\theta c_1 \right] \right]. \quad (24)$$

Now, let $\bar{\theta} = \min\{a/c_1, 1\}$. Then by Lemma 6, we have

$$\limsup_r \mathbb{E} \left[I\{\hat{Z} < 0\} \exp(\theta\hat{\Phi}) \left[-\hat{Z} + (1/2)\theta c_1 \right] \right] \leq C_2(a, c_1),$$

$$\limsup_r \mathbb{E} \left[I\{0 \leq \hat{Z} < a\} \exp(\theta\hat{\Phi}) \left[(1/2)\theta c_1 \right] \right] \leq C_2(a, c_1),$$

for some $C_2(a, c_1) < \infty$. On the other hand $\bar{\theta} = a/c_1$ implies that for any $\theta \in [0, \bar{\theta}]$, $a - (1/2)\theta c_1 > a/2 > 0$. Thus

$$\limsup_r \mathbb{E} \left[I\{\hat{\Phi} < k\} I\{\hat{Z} \geq a\} \exp(\theta\hat{\Phi}) \right] \leq C_3(a, c_1) < \infty,$$

for some $C_3(a, c_1)$. Observe that this bound does not depend on k . Thus fix r_0 large enough so that for all $r \geq r_0$

$$\mathbb{E} \left[I\{\hat{\Phi} < k\} I\{\hat{Z} \geq a\} \exp(\theta\hat{\Phi}) \right] \leq 2C_3(a, c_1).$$

By Monotone Convergence Theorem, taking $k \rightarrow \infty$, we obtain

$$\mathbb{E} \left[I\{\hat{Z} \geq a\} \exp(\theta\hat{\Phi}) \right] \leq 2C_3(a, c_1),$$

namely

$$\limsup_r \mathbb{E} \left[I\{\hat{Z} \geq a\} \exp(\theta\hat{\Phi}) \right] < \infty.$$

Again applying Lemma 6, we have

$$\limsup_r \mathbb{E} \left[I\{\hat{Z} < a\} \exp(\theta \hat{\Phi}) \right] < \infty.$$

Combining, we obtain $\limsup_r \mathbb{E} \left[\exp(\theta \hat{\Phi}) \right] < \infty$.

To complete the proof of the theorem, take any $\theta < \bar{\theta}$ and $p = p(\theta) > 1$ and such that $p\theta < \bar{\theta}$; and then q such that $1/p + 1/q = 1$. Using Holder inequality, we have

$$\mathbb{E} \exp\left(\theta \sum_i (1/\mu_i) \hat{Z}_i^+\right) = \mathbb{E} \left[\exp(\theta \hat{\Phi}) \exp\left(\theta \sum_i (1/\mu_i) \hat{Z}_i^-\right) \right] \leq \left[\mathbb{E} \exp(p\theta \hat{\Phi}) \right]^{1/p} \left[\mathbb{E} \exp(q\theta \sum_i (1/\mu_i) \hat{Z}_i^-) \right]^{1/q}.$$

By Lemma 5 we have

$$\limsup_r \mathbb{E} \left[\exp(q\theta \sum_i (1/\mu_i) \hat{Z}_i^-) \right] < \infty,$$

from which we obtain

$$\limsup_r \mathbb{E} \left[\exp\left(\theta \mu_{\max}^{-1} \sum_i \hat{Z}_i^+\right) \right] < \infty.$$

This proves (4). The proof of Theorem 2(ii) is complete.

5 Proof of Theorem 2(iii)

5.1 Auxiliary results

Lemma 8. *For any $\theta \geq 0$ and any r , the stationary distribution is such that*

$$\mathbb{E} \bar{A} \exp(-\theta \hat{\Phi}) = \sum_{\mathbf{x}} \pi(\mathbf{x}) \bar{A} \exp(-\theta \hat{\Phi}(\mathbf{x})) = 0. \quad (25)$$

Proof. The case $\theta = 0$ is trivial. For the rest of the proof, consider a *fixed pair* of $\theta > 0$ and r . (The chosen constants may depend on θ and r .)

Consider a fixed $c > 0$ and any state \mathbf{x} such that $\hat{\Phi}(\mathbf{x}) \geq c$. Then, it is easy to observe that for some fixed $c_{20} > 0$, $c_{21} > 0$, and any $k \geq c$,

$$\left| \bar{A} \exp(-\theta \hat{\Phi}(\mathbf{x})) \right| \leq (c_{20} \hat{\Phi}(\mathbf{x}) + c_{21}) \exp(-\theta \hat{\Phi}(\mathbf{x})), \quad (26)$$

$$\left| \bar{A} \exp(-\theta \hat{\Phi}_{(k)}(\mathbf{x})) \right| \leq (c_{20} \hat{\Phi}(\mathbf{x}) + c_{21}) \exp(-\theta \hat{\Phi}(\mathbf{x})). \quad (27)$$

Indeed, the total rate of all transitions out of state \mathbf{x} that may change the value of $\hat{\Phi}$ is upper bounded by $c_{20} \hat{\Phi}(\mathbf{x}) + c_{21}$ for some constants $c_{20} > 0$, $c_{21} > 0$. (From $\hat{\Phi} = \sum_i [Z_i - \rho_i r] / (\mu_i \sqrt{r})$ we get an upper bound on Z , linear in $\hat{\Phi}$, which gives the bound on the customer departure rate; the arrival rate is constant.) Also $\hat{\Phi}(\mathbf{y}) \geq \hat{\Phi}(\mathbf{x}) - 1/(\mu_{\min} \sqrt{r}) \geq \hat{\Phi}(\mathbf{x}) - 1/\mu_{\min}$ for any state $\mathbf{y} \in T(\mathbf{x})$. Therefore,

$$\left| \bar{A} \exp(-\theta \hat{\Phi}(\mathbf{x})) \right| \leq (c_{20} \hat{\Phi}(\mathbf{x}) + c_{21}) 2 \exp[-\theta(\hat{\Phi}(\mathbf{x}) - 1/\mu_{\min})],$$

where the factor $2 \exp(\theta/\mu_{\min})$ can be dropped by rechoosing c_{20}, c_{21} . This proves (26), and (27) is proved the same way.

From (26) and (27) we conclude that the values of

$$\left| \bar{A} \exp(-\theta \hat{\Phi}(\mathbf{x})) \right| \quad \text{and} \quad \left| \bar{A} \exp(-\theta \hat{\Phi}_{(k)}(\mathbf{x})) \right|$$

are uniformly bounded for all \mathbf{x} and k . (Recall that the number of states with $\hat{\Phi}(\mathbf{x}) < c$ is finite, for any c .) This means, in particular, that the sum in (25) converges absolutely.

We know that

$$\mathbb{E} \bar{A} \exp(-\theta \hat{\Phi}_{(k)}) = 0, \quad (28)$$

by (2), because $\exp(-\theta \hat{\Phi}_{(k)})$ is constant outside a finite set of states. The LHS of (28) can be written as

$$\mathbb{E} \bar{A} \exp(-\theta \hat{\Phi}_{(k)}) = \sum_{B_1(k)} \pi(\mathbf{x}) \bar{A} \exp(-\theta \hat{\Phi}(\mathbf{x})) + \sum_{B_2(k)} \pi(\mathbf{x}) \bar{A} \exp(-\theta \hat{\Phi}_{(k)}(\mathbf{x})), \quad (29)$$

where $B_1(k)$ is the set of states \mathbf{x} such that \mathbf{x} and all $\mathbf{y} \in T(\mathbf{x})$, are within set $B(k) = \{w \mid \hat{\Phi}(w) < k\}$ (so that $B_1(k) \subseteq B(k)$); $B_2(k)$ is the set of ‘‘boundary’’ states \mathbf{x} for which there exists $\mathbf{y} \in T(\mathbf{x})$ such that either $\mathbf{x} \in B(k)$ or $\mathbf{y} \in B(k)$, but not both. As $k \rightarrow \infty$, the sum over $B_1(k)$ converges to the sum in (25), and the sum over $B_2(k)$ vanishes. This proves (25). \square

Lemma 9. *For arbitrary $b \geq 0$ and $n \in \mathbb{Z}_+$, there exist $\theta_2 = \theta_2(b, n) > 0$ and $C = C(b, n)$ such that for all $\theta \in [0, \theta_2]$,*

$$\limsup_{r \rightarrow \infty} \mathbb{E} \left[I\{\hat{Z} \geq -b\} \left(\sum_i \hat{Z}_i^+ \right)^n \exp(-\theta \hat{\Phi}) \right] \leq C. \quad (30)$$

Proof. The proof is analogous to that of Lemma 6, except instead of relying on the bound (9), which involves Z_i^- and holds when all $\nu_i \leq \mu_i$, we will use bound (4), which involves Z_i^+ and holds for arbitrary $\nu_i \geq 0$. Namely, from $\hat{Z} \geq -b$ we obtain

$$-\hat{\Phi} \leq (1/\mu_{\min} - 1/\mu_{\max}) \sum_i \hat{Z}_i^+ + (1/\mu_{\min})b \quad (31)$$

similarly to (16). Then applying bound (31) to the left-hand side of (30), and then using (4), we obtain (30). \square

5.2 Proof of Theorem 2(iii)

For every state $\mathbf{x} \in \mathcal{X}$ we have

$$\bar{A} \hat{\Phi}(\mathbf{x}) = \frac{1}{\sqrt{r}} \left[\sum_i (\lambda_i r) / \mu_i - \sum_i \Psi_i(\mathbf{x}) - \sum_i \nu_i Q_i(\mathbf{x}) \right] = -\hat{Z}_a(\mathbf{x}) - \sum_i \nu_i \hat{Q}_i(\mathbf{x}),$$

Consider a (candidate) Lyapunov function $\exp(-\theta \hat{\Phi}(\mathbf{x}))$, with $\theta > 0$. Then, analogously to (17) we obtain

$$\begin{aligned} \bar{A} \exp(-\theta \hat{\Phi}(\mathbf{x})) &\leq \exp(-\theta \hat{\Phi}(\mathbf{x})) \left[-\theta \bar{A} \hat{\Phi}(\mathbf{x}) \right. \\ &\quad \left. + (1/2)\theta^2 \left[\sum_i \lambda_i r + (r + a\sqrt{r})\mu_{\max} + \nu_{\max} \sum_i Q_i(\mathbf{x}) \right] (\mu_{\min}^{-1}/\sqrt{r})^2 \exp(\theta \mu_{\min}^{-1}/\sqrt{r}) \right], \end{aligned} \quad (32)$$

where the crude estimate of the maximum total transition rate in the θ^2 term accounts for the transitions due to abandonments as well. In turn, (32) can be rewritten as

$$\bar{A} \exp(-\theta \hat{\Phi}(\mathbf{x})) \leq \exp(-\theta \hat{\Phi}(\mathbf{x})) \left[\theta \hat{Z}_a(\mathbf{x}) + \theta \sum_i \nu_i \hat{Q}_i(\mathbf{x}) + \theta^2 [c_{11} + c_{12} \hat{Z}^+] \right], \quad (33)$$

where $c_{11} > 0$ and $c_{12} > 0$ are constants (independent of r), and the inequality holds for all sufficiently large r . Finally, we fix arbitrary $b > 0$ and rewrite (33) as

$$\bar{A} \exp(-\theta \hat{\Phi}(\mathbf{x})) \leq I\{\hat{Z} < -b\} \exp(-\theta \hat{\Phi}(\mathbf{x})) \left[\theta(-b) + \theta^2 c_{11} \right] + \quad (34)$$

$$I\{\hat{Z} \geq -b\} \exp(-\theta \hat{\Phi}(\mathbf{x})) \left[\theta a + \theta c_{13} \sum_i \hat{Z}_i^+(\mathbf{x}) + \theta^2 [c_{11} + c_{14} \sum_i \hat{Z}_i^+(\mathbf{x})] \right], \quad (35)$$

where $c_{13} > 0$ and $c_{14} > 0$ are constants independent of r .

We now consider our system in steady state. The LHS of (34) has expected value 0 by Lemma 8. For all sufficiently small $\theta > 0$, (35) has expectation bounded uniformly in r by Lemma 9. Further, let us restrict ourselves only to small values of θ such that $b - \theta c_{11} > c_{15}$ for an arbitrary fixed constant $0 < c_{15} < b$. Then, the RHS of (34) has a well defined, strictly negative expected value. (At this point we do not claim that the expectation is finite – that will follow shortly.) We take the expectation of both parts of (34)-(35), cancel θ , move (what is left of) the term (34) to the LHS of the inequality, and divide the inequality by $b - \theta c_{11} > c_{15} > 0$. Letting $r \rightarrow \infty$ and applying Lemma 9 gives us

$$\limsup_r \mathbb{E} \left[I\{\hat{Z} < -b\} \exp(-\theta \hat{\Phi}) \right] < \infty.$$

Again applying Lemma 9, we have

$$\limsup_r \mathbb{E} \left[I\{\hat{Z} \geq -b\} \exp(-\theta \hat{\Phi}) \right] < \infty.$$

Combining, we obtain $\limsup_r \mathbb{E} \left[\exp(-\theta \hat{\Phi}) \right] < \infty$. From this, the required bound (5) for sufficiently small $\theta > 0$, follows by using the identity $\sum_i \hat{Z}_i^- / \mu_i = -\hat{\Phi} + \sum_i \hat{Z}_i^+ / \mu_i$, property (4), and Holder inequality. The proof is complete.

6 Proof of Theorem 4

Statement (i) of the theorem follows from the following result, which is a stronger version of Lemma 5.

Lemma 10. *Suppose $\nu_i \leq \mu_i$ for all i . There exists a sufficiently small $\theta > 0$ such that, for each i ,*

$$\limsup_{r \rightarrow \infty} \mathbb{E} \left[\exp(\theta (\hat{Z}_i^-)^2) \right] < \infty. \quad (36)$$

Given the fact that the vector $(\hat{Z}_i^-, i \in I)$ in Lemma 10 is stochastically dominated by $(\hat{G}_i^-, i \in I)$, where $\hat{G}_i^- = (G_i - \rho_i r) / \sqrt{r}$, and G_i 's are independent r.v. with Poisson distribution with mean $\rho_i r$, this lemma in turn is a corollary of the following asymptotic property of Poisson distributions.

Lemma 11. *Consider a family of Poisson distributed random variables $H(p)$ with mean $p > 0$. Denote*

$$h_n = h_n(p) = \mathbb{P}\{H(p) = n\} = e^{-p} p^n / n!, n \in \mathbb{Z}_+.$$

Then, there exists a constant $C > 0$, such that for all sufficiently large p and all $n \leq p$,

$$h_n \leq C \frac{1}{\sqrt{p}} e^{-(n-p)^2 / (2p)}. \quad (37)$$

Consequently, for any $0 \leq \theta < 1/2$,

$$\limsup_{p \rightarrow \infty} \mathbb{E} e^{\theta [(H(p)-p)]^2 / p} < \infty. \quad (38)$$

Proof. Denote $m = \lfloor p \rfloor$. Using Stirling formula for the factorial in $h_m = e^{-p} p^m / m!$, we directly verify that

$$\lim_{p \rightarrow \infty} h_m \left[\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{p}} \right]^{-1} = 1,$$

in particular, for some constant $C_1 > 0$ and all large p

$$h_m \leq C_1 \frac{1}{\sqrt{p}}. \quad (39)$$

Therefore, to show (37), it will suffice to show the existence of $C_2 > 0$ such that for all large p and any $n \leq m$, we have

$$h_n / h_m \leq C_2 e^{-(n-p)^2 / (2p)}. \quad (40)$$

We only need to verify (40) for $n < m$, because $(m-p)^2 / (2p) \rightarrow 0$ as $p \rightarrow \infty$, so that the case $n = m$ can be covered by rechoosing (increasing) C_2 if necessary. We have

$$\log(h_n / h_m) = \log \prod_{j=n+1}^m \frac{j}{p} = \sum_{j=n+1}^m \log\left(1 - \frac{p-j}{p}\right) \leq - \sum_{j=n+1}^m \frac{p-j}{p} = -\frac{1}{2p}(m-n)(p-m+p-n-1).$$

Note that $\epsilon = p - m \in [0, 1)$. Substituting $m = p - \epsilon$ in the RHS of the above inequality, we see that it is

$$-\frac{1}{2p}(p-n)^2 + \frac{1}{2p}(p-n+\epsilon(1-\epsilon)) \leq -\frac{1}{2p}(p-n)^2 + 1$$

for all $p \geq 1$. We proved (37).

To show (38), we can write

$$\begin{aligned} \mathbb{E} e^{\theta[(H(p)-p)^{-1}]^2/p} &\leq 1 + \sum_{n < p} h_n e^{\theta(n-p)^2/p} \leq \\ &1 + \sum_{n < p-1} C \frac{1}{\sqrt{p}} e^{(\theta-1/2)(n-p)^2/p} + C \frac{1}{\sqrt{p}}. \end{aligned}$$

It remains to observe that

$$\sum_{n < p-1} \frac{1}{\sqrt{p}} e^{(\theta-1/2)(n-p)^2/p} \leq \int_{-\infty}^0 e^{-(1/2-\theta)\xi^2} d\xi.$$

□

Thus, statement (i) of the theorem is proved. In view of Lemma 7, to prove statement (ii), it suffices to consider the case $\nu_i \leq \mu_i$ for all i . Then the following proposition will imply statement (ii) almost immediately.

Proposition 12. *Suppose, $0 < \nu_i \leq \mu_i$ for all i . There exists $\bar{\theta} > 0$ and $c_0 > 0$ such that for all $\theta \in [0, \bar{\theta}]$ and k*

$$\limsup_{r \rightarrow \infty} \mathbb{E} \left[I\{\hat{\Phi} \leq k\} \exp(\theta \hat{\Phi}^2) \right] \leq c_0, \quad (41)$$

$$\limsup_{r \rightarrow \infty} \mathbb{E} \left[I\{\hat{\Phi} \leq k\} \exp\left(\theta \left(\sum_i \hat{Z}_i^+ / \mu_i\right)^2\right) \right] \leq c_0. \quad (42)$$

Proof of Proposition 12. Property (42) follows from Lemma 10, (41) and application of Cauchy-Schwartz inequality because

$$\left(\sum_i \hat{Z}_i^+ / \mu_i\right)^2 = (\hat{\Phi} + \sum_i \hat{Z}_i^- / \mu_i)^2 \leq 2\hat{\Phi}^2 + 2|I| \sum_i (\hat{Z}_i^- / \mu_i)^2.$$

Thus, the main bulk of the proof is devoted to (41).

We first write the expression and derive bounds for $\bar{A}\hat{\Phi}$. We have (using notation $\Psi = \sum_i \Psi_i$) for every state \mathbf{x}

$$\begin{aligned}\bar{A}\hat{\Phi}(\mathbf{x}) &= \frac{1}{\sqrt{r}}[\sum_i (\lambda_i r)/\mu_i - \sum_i \Psi_i(\mathbf{x}) - \sum_i (Z_i(\mathbf{x}) - \Psi_i(\mathbf{x}))\nu_i/\mu_i] \\ &\leq \frac{1}{\sqrt{r}}[r - \Psi(\mathbf{x}) - \nu_{\min}/\mu_{\max}(Z(\mathbf{x}) - \Psi(\mathbf{x}))] \\ &= -\min\{\hat{Z}(\mathbf{x}), a\} - \nu_{\min}/\mu_{\max}(\hat{Z}(\mathbf{x}) - a)^+.\end{aligned}\tag{43}$$

Now we have

$$\hat{Z}(\mathbf{x}) = \sum_i \hat{Z}_i(\mathbf{x}) = \sum_i \hat{Z}_i^+(\mathbf{x}) - \sum_i \hat{Z}_i^-(\mathbf{x}) \geq \mu_{\min}\hat{\Phi}^+(\mathbf{x}) - \sum_i \hat{Z}_i^-(\mathbf{x}).$$

Since $(x - a)^+$ is a monotone function, we obtain a bound

$$\begin{aligned}\bar{A}\hat{\Phi}(\mathbf{x}) &\leq -\min\{\hat{Z}(\mathbf{x}), a\} - (\nu_{\min}\mu_{\min}/\mu_{\max})\hat{\Phi}^+(\mathbf{x}) + (\nu_{\min}/\mu_{\max})\sum_i \hat{Z}_i^-(\mathbf{x}) + (\nu_{\min}/\mu_{\max})a \\ &\triangleq -\min\{\hat{Z}(\mathbf{x}), a\} - c_1\hat{\Phi}^+ + c_2\sum_i \hat{Z}_i^-(\mathbf{x}) + c_3.\end{aligned}\tag{44}$$

Similarly to (44) we can also obtain a lower bound on $\bar{A}\hat{\Phi}$. Namely, from (43) we get

$$\bar{A}\hat{\Phi}(\mathbf{x}) \geq -\min\{\hat{Z}(\mathbf{x}), a\} - \nu_{\max}/\mu_{\min}(\hat{Z}(\mathbf{x}) - a)^+,$$

which, using $\hat{Z}(\mathbf{x}) \leq \sum_i \hat{Z}_i^+(\mathbf{x})$, implies

$$\bar{A}\hat{\Phi}(\mathbf{x}) \geq -\min\{\hat{Z}(\mathbf{x}), a\} - c_{25}\sum_i \hat{Z}_i^+(\mathbf{x}),\tag{45}$$

where $c_{25} = \nu_{\max}/\mu_{\min}$.

Consider a (candidate) Lyapunov function $\exp(\theta\hat{\Phi}^2(\mathbf{x}))$ with $0 < \theta \leq \bar{\theta}$ where $\bar{\theta}$ will be implicitly defined below. (Below we use the phrase “for sufficiently small $\theta > 0$ ” to mean “for all $\theta \in (0, \bar{\theta}]$, with $\bar{\theta}$ rechosen to be smaller, if necessary.”) Without the loss of generality we assume $\bar{\theta} \leq 1$. We have for $h(x) = \exp(\theta x^2)$ that $dh/dx = 2\theta x \exp(\theta x^2)$ and $d^2h/dx^2 = 2\theta \exp(\theta x^2) + 4\theta^2 x^2 \exp(\theta x^2)$. We use Taylor expansion

$$\exp(\theta y^2) = \exp(\theta x^2) \left(1 + 2\theta x(y - x) + \frac{1}{2}(2\theta + 4\theta^2 z^2)(y - x)^2 \exp(\theta z^2 - \theta x^2) \right)$$

for some z satisfying $|z - x| \leq |y - x|$. We have $z^2 - x^2 = (z - x)(z + x) \leq 2|y - x|(|y| + |x|)$. Then

$$\exp(\theta y^2) \leq \exp(\theta x^2) (1 + 2\theta x(y - x) + (\theta + 2\theta^2(|x| + |y - x|)^2)(y - x)^2 \exp(\theta 2|y - x|(|y| + |x|)))$$

Let $\hat{\Phi}_{(k)} = \min(\hat{\Phi}, k)$ and let us apply the above inequalities for $x = \hat{\Phi}_{(k)}(\mathbf{x})$ and $y = \hat{\Phi}_{(k)}(\mathbf{y})$, where $\mathbf{y} \in T(\mathbf{x})$. We have $|\hat{\Phi}_{(k)}(\mathbf{y}) - \hat{\Phi}_{(k)}(\mathbf{x})| \leq \mu_{\min}^{-1}/\sqrt{r}$ implying the same for $\hat{\Phi}_{(k)}$. Then

$$\max\{\hat{\Phi}_{(k)}(\mathbf{y}), \hat{\Phi}_{(k)}(\mathbf{x})\} \leq k,$$

and

$$\min\{\hat{\Phi}_{(k)}(\mathbf{y}), \hat{\Phi}_{(k)}(\mathbf{x})\} \geq -\hat{\Phi}^-(\mathbf{x}) - \mu_{\min}^{-1}/\sqrt{r}.$$

This implies

$$2|\hat{\Phi}_{(k)}(\mathbf{y}) - \hat{\Phi}_{(k)}(\mathbf{x})|(|\hat{\Phi}_{(k)}(\mathbf{y})| + |\hat{\Phi}_{(k)}(\mathbf{x})|) \leq 4[\mu_{\min}^{-2}/r + \mu_{\min}^{-1}k/\sqrt{r} + \mu_{\min}^{-1}\hat{\Phi}^-(\mathbf{x})/\sqrt{r}].$$

For sufficiently large r which depends on k and model parameters, we obtain even a cruder upper bound

$$2|\hat{\Phi}_{(k)}(\mathbf{y}) - \hat{\Phi}_{(k)}(\mathbf{x})|(|\hat{\Phi}_{(k)}(\mathbf{y})| + |\hat{\Phi}_{(k)}(\mathbf{x})|) \leq \hat{\Phi}^-(\mathbf{x}) + 1$$

giving

$$\exp(\theta 2|\hat{\Phi}_{(k)}(\mathbf{y}) - \hat{\Phi}_{(k)}(\mathbf{x})|(|\hat{\Phi}_{(k)}(\mathbf{y})| + |\hat{\Phi}_{(k)}(\mathbf{x})|)) \leq \exp(\theta \hat{\Phi}^-(\mathbf{x}) + \theta).$$

Note that, given k , the value of $\hat{\Phi}_{(k)}(\mathbf{x})$ can change after a direct transition only when $\sum_i \hat{Z}_i(\mathbf{x})/\mu_i \leq k + 1/\mu_{\min}$, implying $Z \leq r + \mu_{\max}(k + 1/\mu_{\min})\sqrt{r}$. Therefore, uniformly on all states \mathbf{x} , the total rate of all transitions that can possibly change the value of $\hat{\Phi}_{(k)}(\mathbf{x})$ is upper bounded by

$$r \sum_i \lambda_i + (r + a\sqrt{r})\mu_{\max} + \sqrt{r}\mu_{\max}(k + 1/\mu_{\min}) \sum_i \nu_i \leq c_4 r,$$

for the appropriate choice of positive constant c_4 . Also,

$$(\hat{\Phi}_{(k)}(\mathbf{y}) - \hat{\Phi}_{(k)}(\mathbf{x}))^2 \leq 1/(\mu_{\min}^2 r).$$

Therefore, the product of the rate of transitions (which can change $\hat{\Phi}_{(k)}$) and the value of $(\hat{\Phi}_{(k)}(\mathbf{y}) - \hat{\Phi}_{(k)}(\mathbf{x}))^2$ is upper bounded by c_4/μ_{\min}^2 . Assembling all the above estimates, we obtain

$$\begin{aligned} A \exp(\theta \hat{\Phi}_{(k)}^2(\mathbf{x})) &\leq 2\theta \hat{\Phi}_{(k)}(\mathbf{x}) [A \hat{\Phi}_{(k)}(\mathbf{x})] \exp(\theta \hat{\Phi}_{(k)}^2(\mathbf{x})) \\ &\quad + \theta c_4 \mu_{\min}^{-2} \exp(\theta \hat{\Phi}^-(\mathbf{x}) + \theta) \exp(\theta \hat{\Phi}_{(k)}^2(\mathbf{x})) \\ &\quad + 2\theta^2 (|\hat{\Phi}_{(k)}(\mathbf{x})| + \mu_{\min}^{-1} r^{-1/2})^2 \exp(\theta \hat{\Phi}^-(\mathbf{x}) + \theta) \exp(\theta \hat{\Phi}_{(k)}^2(\mathbf{x})) \\ &\leq 2\theta \hat{\Phi}_{(k)}(\mathbf{x}) [A \hat{\Phi}_{(k)}(\mathbf{x})] \exp(\theta \hat{\Phi}_{(k)}^2(\mathbf{x})) \\ &\quad + \theta c_5 \exp(\theta \hat{\Phi}^-(\mathbf{x})) \exp(\theta \hat{\Phi}_{(k)}^2(\mathbf{x})) \\ &\quad + c_6 \theta^2 (|\hat{\Phi}_{(k)}(\mathbf{x})| + \mu_{\min}^{-1} r^{-1/2})^2 \exp(\theta \hat{\Phi}^-(\mathbf{x})) \exp(\theta \hat{\Phi}_{(k)}^2(\mathbf{x})) \end{aligned}$$

where we have incorporated e^θ into constants c_5 and c_6 . This is possible by our assumption $\bar{\theta} \leq 1$. Furthermore, for large enough r we have $\mu_{\min}^{-1} r^{-1/2} < 1$. Thus $(|\hat{\Phi}_{(k)}(\mathbf{x})| + \mu_{\min}^{-1} r^{-1/2})^2 \leq 2\hat{\Phi}_{(k)}^2(\mathbf{x}) + 2$, and for the appropriate choice of constants c_7 and c_8 , we obtain

$$\begin{aligned} A \exp(\theta \hat{\Phi}_{(k)}^2(\mathbf{x})) &\leq 2\theta \hat{\Phi}_{(k)}(\mathbf{x}) [A \hat{\Phi}_{(k)}(\mathbf{x})] \exp(\theta \hat{\Phi}_{(k)}^2(\mathbf{x})) \\ &\quad + \theta c_7 \exp(\theta \hat{\Phi}^-(\mathbf{x})) \exp(\theta \hat{\Phi}_{(k)}^2(\mathbf{x})) \\ &\quad + c_8 \theta^2 \hat{\Phi}_{(k)}^2(\mathbf{x}) \exp(\theta \hat{\Phi}^-(\mathbf{x})) \exp(\theta \hat{\Phi}_{(k)}^2(\mathbf{x})) \end{aligned}$$

Now we will use appropriate bounds on $A\hat{\Phi}_{(k)}(\mathbf{x})$. Observe that if the state \mathbf{x} is such that $\hat{\Phi}(\mathbf{x}) > k$ then the left hand side is at most zero. If \mathbf{x} is such that $0 \leq \hat{\Phi}(\mathbf{x}) \leq k$, then we have $A\hat{\Phi}_{(k)}(\mathbf{x}) \leq \bar{A}\hat{\Phi}(\mathbf{x})$ and can use upper bound (44) on $\bar{A}\hat{\Phi}(\mathbf{x})$. Also, for all large enough r , $\hat{\Phi}(\mathbf{x}) < 0$ implies $A\hat{\Phi}_{(k)}(\mathbf{x}) = \bar{A}\hat{\Phi}(\mathbf{x})$ and we can use lower bound (45) on $\bar{A}\hat{\Phi}(\mathbf{x})$.

Therefore, we can write

$$\begin{aligned}
A \exp(\theta \hat{\Phi}_{(k)}^2(\mathbf{x})) &\leq -2\theta \hat{\Phi}(\mathbf{x}) I\{\hat{\Phi}(\mathbf{x}) \leq k\} \exp(\theta \hat{\Phi}^2(\mathbf{x})) \min\{\hat{Z}(\mathbf{x}), a\} \\
&\quad - 2c_1 \theta \hat{\Phi}(\mathbf{x}) \hat{\Phi}^+(\mathbf{x}) I\{0 \leq \hat{\Phi}(\mathbf{x}) \leq k\} \exp(\theta \hat{\Phi}^2(\mathbf{x})) \\
&\quad + 2c_2 \theta \hat{\Phi}(\mathbf{x}) I\{0 \leq \hat{\Phi}(\mathbf{x}) \leq k\} \exp(\theta \hat{\Phi}^2(\mathbf{x})) \sum_i \hat{Z}_i^- \\
&\quad + 2c_3 \theta \hat{\Phi}(\mathbf{x}) I\{0 \leq \hat{\Phi}(\mathbf{x}) \leq k\} \exp(\theta \hat{\Phi}^2(\mathbf{x})) \\
&\quad - 2c_{25} \theta \hat{\Phi}(\mathbf{x}) I\{\hat{\Phi}(\mathbf{x}) < 0\} \exp(\theta \hat{\Phi}^2(\mathbf{x})) \sum_i \hat{Z}_i^+ \\
&\quad + c_7 \theta I\{\hat{\Phi}(\mathbf{x}) \leq k\} \exp(\theta \hat{\Phi}^-(\mathbf{x})) \exp(\theta \hat{\Phi}^2(\mathbf{x})) \\
&\quad + c_8 \theta^2 I\{\hat{\Phi}(\mathbf{x}) \leq k\} \hat{\Phi}^2(\mathbf{x}) \exp(\theta \hat{\Phi}^-(\mathbf{x})) \exp(\theta \hat{\Phi}^2(\mathbf{x})),
\end{aligned}$$

where we switched from $\hat{\Phi}_{(k)}(\mathbf{x})$ to $\hat{\Phi}(\mathbf{x})$ since they are equal when $\hat{\Phi} \leq k$. We take the expectation of both sides with respect to the stationary distribution of the process. Since the expectation of the LHS is zero, and we can get rid of positive factor θ , we obtain

$$\begin{aligned}
2\mathbb{E} \left[\hat{\Phi} I\{\hat{\Phi} \leq k\} \exp(\theta \hat{\Phi}^2) \min\{\hat{Z}, a\} \right] &+ 2c_1 \mathbb{E} \left[\hat{\Phi} \hat{\Phi}^+ I\{0 \leq \hat{\Phi} \leq k\} \exp(\theta \hat{\Phi}^2) \right] \\
&\leq 2c_2 \mathbb{E} \left[\hat{\Phi}^+ I\{0 \leq \hat{\Phi} \leq k\} \exp(\theta (\hat{\Phi}^+)^2) \sum_i \hat{Z}_i^- \right] \quad (46) \\
&\quad + 2c_3 \mathbb{E} \left[\hat{\Phi}^+ I\{0 \leq \hat{\Phi} \leq k\} \exp(\theta (\hat{\Phi}^+)^2) \right] \\
&\quad + 2c_{25} \mathbb{E} \left[\hat{\Phi}^- I\{\hat{\Phi} < 0\} \exp(\theta (\hat{\Phi}^+)^2) \sum_i \hat{Z}_i^+ \right] \\
&\quad + c_7 \theta \mathbb{E} \left[I\{\hat{\Phi} \leq k\} \exp(\theta \hat{\Phi}^-) \exp(\theta \hat{\Phi}^2) \right] \\
&\quad + c_8 \theta \mathbb{E} \left[I\{\hat{\Phi} \leq k\} \hat{\Phi}^2 \exp(\theta \hat{\Phi}^-) \exp(\theta \hat{\Phi}^2) \right],
\end{aligned}$$

where in the first two terms after \leq , we replaced $\hat{\Phi}$ by $\hat{\Phi}^+$ because they are equal when $\hat{\Phi} \geq 0$; similarly, in the third term after \leq , we replaced $-\hat{\Phi}$ and $\hat{\Phi}^2$ by $\hat{\Phi}^-$ and $(\hat{\Phi}^-)^2$, respectively, using condition $\hat{\Phi} < 0$.

Let us analyze term by term and obtain some bounds that are uniform in r and k . First consider

$$2\mathbb{E} \left[\hat{\Phi} I\{\hat{\Phi} \leq k\} \exp(\theta \hat{\Phi}^2) \min\{\hat{Z}, a\} \right].$$

When $\hat{\Phi}$ and \hat{Z} have the same sign, the term is non-negative and we can drop it from the inequality. When $\hat{\Phi} \leq 0$ and $\hat{Z} \geq 0$ we have

$$|2\hat{\Phi} I\{\hat{\Phi} \leq k\} \exp(\theta \hat{\Phi}^2) \min\{\hat{Z}, a\}| \leq 2a \hat{\Phi}^- \exp(\theta \hat{\Phi}^2) \leq 2a \mu_{\max} \left(\sum_i \hat{Z}_i^- \right) \exp(\theta \mu_{\min}^{-2} \left(\sum_i \hat{Z}_i^- \right)^2).$$

For sufficiently small $\theta > 0$, applying Lemma 10, we have

$$\limsup_{r \rightarrow \infty} 2\mathbb{E} \left[I\{\hat{\Phi} \leq 0, \hat{Z} \geq 0\} \left| \hat{\Phi} I\{\hat{\Phi} \leq k\} \exp(\theta \hat{\Phi}^2) \min\{\hat{Z}, a\} \right| \right] \leq c_9,$$

for some $0 < c_9 < \infty$ (independent of k). Finally, when $\hat{\Phi} \geq 0, \hat{Z} \leq 0$ we have

$$2|\hat{\Phi} I\{\hat{\Phi} \leq k\} \exp(\theta \hat{\Phi}^2) \min\{\hat{Z}, a\}| \leq 2\hat{\Phi}^+ I\{\hat{\Phi} \leq k\} \exp(\theta (\hat{\Phi}^+)^2) \sum_i \hat{Z}_i^-.$$

Observe that we already have the expectation of such a term on the right hand side of our inequality (46), up to a constant factor.

For the second term in the inequality (46), we have

$$2c_1\mathbb{E}\left[\hat{\Phi}\hat{\Phi}^+I\{\hat{\Phi}\leq k\}\exp(\theta\hat{\Phi}^2)\right]=2c_1\mathbb{E}\left[(\hat{\Phi}^+)^2I\{\hat{\Phi}\leq k\}\exp(\theta(\hat{\Phi}^+)^2)\right]=g, \quad (47)$$

where $g = g(r, k, \theta)$ is simply a notation for the RHS, for the ease of reference.

We leave the first term after inequality (46) untouched for now. The second term after inequality (46) is upper bounded by $g/4 + c_{10}$ for some constant $c_{10} > 0$, by considering the cases $\hat{\Phi} < c$ and $\hat{\Phi} \geq c$ separately, for a sufficiently large fixed $c > 0$. For the third term after inequality (46), we use the fact that for some fixed $c_{26} > 0$, condition $\hat{\Phi} < 0$ implies $\sum_i \hat{Z}_i^+ \leq c_{26} \sum_i \hat{Z}_i^-$, and therefore, by Lemma 10, this term is uniformly bounded for all r and all sufficiently small $\theta > 0$.

Next consider

$$c_7\mathbb{E}\left[I\{\hat{\Phi}\leq k\}\exp(\theta\hat{\Phi}^-)\exp(\theta\hat{\Phi}^2)\right].$$

The expectation on the event $\{\hat{\Phi} \leq 0\}$ is bounded by $c_7\mathbb{E}[\exp(\theta\hat{\Phi}^-)\exp(\theta(\hat{\Phi}^-)^2)]$ which, by Lemma 10, is bounded by some $c_{11} > 0$ and sufficiently small $\theta > 0$, uniformly in r . The expectation on the event $\{\hat{\Phi} > 0\}$ is bounded by $c_7\mathbb{E}\left[I\{\hat{\Phi}\leq k\}\exp(\theta(\hat{\Phi}^+)^2)\right]$ which is upper bounded by $g/4 + c_{12}$ for some constant $c_{12} > 0$, again by considering the cases $\hat{\Phi} < c$ and $\hat{\Phi} \geq c$ separately, for a sufficiently large fixed $c > 0$.

Finally, for the last term in (46)

$$c_8\theta\mathbb{E}\left[I\{\hat{\Phi}\leq k\}\hat{\Phi}^2\exp(\theta\hat{\Phi}^-)\exp(\theta\hat{\Phi}^2)\right],$$

the expectation on the event $\{\hat{\Phi} < 0\}$ is uniformly bounded by some $c_{13}\theta$, for sufficiently small $\theta > 0$, by applying Lemma 10. The expectation on the event $\{\hat{\Phi} \geq 0\}$ is bounded by

$$c_8\theta\mathbb{E}\left[I\{\hat{\Phi}^+\leq k\}(\hat{\Phi}^+)^2\exp(\theta(\hat{\Phi}^+)^2)\right]$$

For sufficiently small $\theta > 0$, to ensure $c_8\theta < (1/2)c_1$, this term is upper bounded by $g/4$. Thus we can rewrite (46) as

$$c_{21}\mathbb{E}\left[(\hat{\Phi}^+)^2I\{\hat{\Phi}\leq k\}\exp(\theta(\hat{\Phi}^+)^2)\right]\leq c_{22}\mathbb{E}\left[\hat{\Phi}^+I\{\hat{\Phi}\leq k\}\exp(\theta(\hat{\Phi}^+)^2)\sum_i\hat{Z}_i^-\right]+c_{14}, \quad (48)$$

for some positive c_{21}, c_{22}, c_{14} . We analyze the term on the right-hand side. Fix $C > 0$. Then

$$\begin{aligned} \hat{\Phi}^+I\{\hat{\Phi}\leq k\}\exp(\theta(\hat{\Phi}^+)^2)\sum_i\hat{Z}_i^- &= \hat{\Phi}^+I\{C\sum_i\hat{Z}_i^-\leq\hat{\Phi}\leq k\}\exp(\theta(\hat{\Phi}^+)^2)\sum_i\hat{Z}_i^- \\ &\quad + \hat{\Phi}^+I\{\hat{\Phi}<\min(k,C\sum_i\hat{Z}_i^-)\}\exp(\theta\hat{\Phi}^2)\sum_i\hat{Z}_i^- \\ &\leq (1/C)(\hat{\Phi}^+)^2I\{\hat{\Phi}\leq k\}\exp(\theta(\hat{\Phi}^+)^2) \\ &\quad + C(\sum_i\hat{Z}_i^-)^2\exp(\theta C^2(\sum_i\hat{Z}_i^-)^2) \end{aligned}$$

The second summand is bounded in expectation (by Lemma 10) for sufficiently small $\theta > 0$. Consider constant C such that $c_{21} - c_{22}/C = c_{21}/2$ (that is $C = 2c_{22}/c_{21}$). By the choice of C , if we apply the above estimates to the RHS of (48), we obtain

$$(c_{21}/2)\mathbb{E}\left[(\hat{\Phi}^+)^2I\{\hat{\Phi}\leq k\}\exp(\theta(\hat{\Phi}^+)^2)\right]\leq c_{24}$$

for some $c_{24} > 0$. We have identified a constant c_{15} such that for every k

$$\limsup_{r \rightarrow \infty} \mathbb{E} \left[(\hat{\Phi}^+)^2 I\{\hat{\Phi} \leq k\} \exp(\theta(\hat{\Phi}^+)^2) \right] \leq c_{15} < \infty.$$

Since, again by Lemma 10, for sufficiently small $\theta > 0$, we also have for some c_{16}

$$\limsup_{r \rightarrow \infty} \mathbb{E} \left[(\hat{\Phi}^-)^2 I\{\hat{\Phi} \leq k\} \exp(\theta(\hat{\Phi}^-)^2) \right] \leq c_{16} < \infty,$$

we obtain for $c_0 \triangleq c_{15} + c_{16}$

$$\limsup_{r \rightarrow \infty} \mathbb{E} \left[\hat{\Phi}^2 I\{\hat{\Phi} \leq k\} \exp(\theta(\hat{\Phi}^2)) \right] \leq c_0 < \infty.$$

This implies (41) and the proof of Proposition 12 is complete. \square

We now complete the proof of Theorem 4(ii). To show (7), we use the argument similar to that in the proof of Corollary 3. Namely, we fix a sufficiently small $\theta > 0$ for which the statement of Proposition 12 holds. The weak convergence of $\hat{\Phi}^r(\infty)$ to $\hat{\Phi}(\infty)$ implies that, for any fixed k ,

$$\mathbb{E} \left[I\{|\hat{\Phi}(\infty)| \leq k\} \exp(\theta(\hat{\Phi}(\infty))^2) \right] \leq \limsup_{r \rightarrow \infty} \mathbb{E} \left[I\{\hat{\Phi}^r(\infty) \leq 2k\} \exp(\theta(\hat{\Phi}^r(\infty))^2) \right] \leq c_0.$$

Then, by Monotone Convergence Theorem,

$$\mathbb{E} \left[\exp(\theta(\hat{\Phi}(\infty))^2) \right] \leq c_0.$$

Finally, we use

$$\mu_{max}^{-2} (\hat{Z}^+(\infty))^2 \leq \left(\sum_i \hat{Z}_i^+(\infty) / \mu_i \right)^2 = (\hat{\Phi}(\infty) + \sum_i \hat{Z}_i^-(\infty) / \mu_i)^2 \leq 2(\hat{\Phi}(\infty))^2 + 2|I| \sum_i (\hat{Z}_i^-(\infty) / \mu_i)^2,$$

and apply of Cauchy-Schwartz inequality with θ rechosen to be still smaller, if necessary, so that Proposition 12 and Lemma 10 can be used.

References

- [AAM07] Z. Aksin, M. Armony, and V. Mehrotra, *The modern call-center: A multi-disciplinary perspective on operations management research*, Production and Operations Management, Special Issue on Service Operations in honor of John Buzacott (ed. G. Shanthikumar and D. Yao) **16** (2007), no. 6, 655–688.
- [AG] B. Ata and I. Gurvich, *On optimality gaps in the Halfin-Whitt regime*, Preprint.
- [AMR04a] R. Atar, A. Mandelbaum, and M. Reiman, *A Brownian control problem for a simple queueing system in the Halfin-Whitt regime*, Systems and Control Letters **51** (2004), no. 3-4, 269–275.
- [AMR04b] ———, *Scheduling a multi-class queue with many exponential servers*, Ann. Appl. Prob. **14** (2004), 1084–1134.
- [BGM⁺05] 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 **100** (2005), 36–50.
- [BL09] A. Budhiraja and C. Lee, *Stationary distribution convergence for generalized Jackson networks in heavy traffic*, Math. Oper. Res. **34** (2009), 45–56.

- [DH10] J. G. Dai and S. He, *Many-server queues with customer abandonment: numerical analysis of their diffusion models*, *Annals of Applied Probability* **20** (2010), 1854–1890.
- [DT10] J. G. Dai and T. Tezcan, *Dynamic control of n -systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic*, *Operations Research* **58** (2010), 94–110.
- [FPG08] R. R. Mazumdar F. Piera and F. Guillemin, *Boundary behavior and product-form stationary distributions of jump-diffusions in the orthant with state dependent reflections*, *Advances in Applied Probability* **40** (2008), 529–547.
- [GG11] D. Gamarnik and D. Goldberg, *Steady-state GI/GI/ n queue in the Halfin-Whitt regime*, Preprint at <http://arxiv.org/abs/1103.1709> (2011).
- [GM08] D. Gamarnik and P. Momčilović, *Steady-state analysis of a multi-server queue in the Halfin-Whitt regime*, *Advances in Applied Probability* **40** (2008), 548–577.
- [GMR02] O. Garnett, A. Mandelbaum, and M. Reiman, *Designing a call center with impatient customers*, *Manufacturing and Service Operations Management* **4** (2002), 208–227.
- [GZ06] D. Gamarnik and A. Zeevi, *Validity of heavy traffic steady-state approximations in open queueing networks*, *Ann. Appl. Prob.* **16** (2006), no. 1, 56–90.
- [HW81] S. Halfin and W. Whitt, *Heavy-traffic limits for queues with many exponential servers*, *Operations Research* **29** (1981), no. 3, 567–588.
- [JMM04] P. Jelenković, A. Mandelbaum, and P. Momčilović, *Heavy traffic limits for queues with many deterministic servers*, *Queueing Systems* **47** (2004), 53–69.
- [Kat10] T. Katsuda, *State-space collapse in stationarity and its application to a multiclass single-server queue in heavy traffic*, *Queueing Systems* **65** (2010), 237273.
- [KR] W. Kang and K. Ramanan, *Asymptotic approximations for the stationary distributions of many-server queues with abandonment*, Preprint.
- [KR01] H. Kaspi and K. Ramanan, *Law of large numbers limits for many-server queues*, *Ann. of Appl. Prob.* **21** (2001), 33–114.
- [MM] A. Mandelbaum and P. Momcilovic, *Queues with many servers and impatient customers*, Preprint.
- [MM05] ———, *Queues with many servers: The virtual waiting-time process in the QED regime*, *Mathematics of Operations Research* **33** (2005), 561586.
- [NGM03] G. Koole N. Gans and A. Mandelbaum, *Telephone call centers: Tutorial, review, and research prospects*, *Manufacturing & Service Operations Management* **5** (2003), 79–141.
- [Ree09] J. E Reed, *The G/GI/ N queue in the Halfin-Whitt regime*, *Annals of Applied Probability* **19** (2009), 2211–2269.
- [RP00] M. I. Reiman and A. A. Puhalskii, *The multiclass GI/PH/ N queue in the Halfin-Whitt regime*, *Advances in Applied Probability* **32** (2000), 564–595.
- [ST10] A.L. Stolyar and T. Tezcan, *Control of systems with flexible multi-server pools: A shadow routing approach*, *Queueing Systems* **66** (2010), 1–51.