http://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 46, No. 2, May 2021, pp. 772–796
ISSN 0364-765X (print), ISSN 1526-5471 (online)

On Uniform Exponential Ergodicity of Markovian Multiclass Many-Server Queues in the Halfin–Whitt Regime

Ari Arapostathis,[a] Hassan Hmedi,[a] Guodong Pang[b]

[a] Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, Texas 78712; [b] Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, College of Engineering, Pennsylvania State University, University Park, Pennsylvania 16802

Contact: ari@utexas.edu, https://orcid.org/0000-0003-2207-357X (AA); hmedi@utexas.edu, https://orcid.org/0000-0003-0104-8279 (HH); gup3@psu.edu, https://orcid.org/0000-0001-6468-0429 (GP)

Received: December 11, 2018
Revised: May 7, 2019; September 28, 2019; November 21, 2019
Accepted: January 12, 2020
Published Online in Articles in Advance: February 3, 2021

MSC2000 Subject Classification: Primary: 90B22, 60K25, 90B15

https://doi.org/10.1287/moor.2020.1087

Abstract. We study ergodic properties of Markovian multiclass many-server queues that are uniform over scheduling policies and the size of the system. The system is heavily loaded in the Halfin–Whitt regime, and the scheduling policies are work conserving and preemptive. We provide a unified approach via a Lyapunov function method that establishes Foster–Lyapunov equations for both the limiting diffusion and the prelimit diffusion-scaled queuing processes simultaneously. We first study the limiting controlled diffusion and show that if the spare capacity (safety staffing) parameter is positive, the diffusion is exponentially ergodic uniformly over all stationary Markov controls, and the invariant probability measures have uniform exponential tails. This result is sharp because when there is no abandonment and the spare capacity parameter is negative, the controlled diffusion is transient under any Markov control. In addition, we show that if all the abandonment rates are positive, the invariant probability measures have sub-Gaussian tails regardless whether the spare capacity parameter is positive or negative. Using these results, we proceed to establish the corresponding ergodic properties for the diffusion-scaled queuing processes. In addition to providing a simpler proof of previous results in Gamarnik and Stolyar [Gamarnik D, Stolyar AL (2012) Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime: asymptotics of the stationary distribution. *Queueing Systems* 71(1–2):25–51], we extend these results to multiclass models with renewal arrival processes, albeit under the assumption that the mean residual life functions are bounded. For the Markovian model with Poisson arrivals, we obtain stronger results and show that the convergence to the stationary distribution is at an exponential rate uniformly over all work-conserving stationary Markov scheduling policies.

Funding: This work was supported by the National Science Foundation [Grants 1715210, 1635410, and 1715875], the Office of Naval Research [Grant N00014-16-1-2956], and the Army Research Office [Grant W911NF-17-1-0019].

Keywords: multiclass many-server queues • Halfin–Whitt (QED) regime • uniform exponential ergodicity • diffusion scaling

1. Introduction

Multiclass many-server queues in the Halfin–Whitt (H-W) regime have been extensively studied as a useful model for large-scale service systems. In this paper, we focus on ergodic properties of such multiclass queuing models. The ergodic properties of these systems have been the subject of great interest in applied probability (Dai and He [15], Gamarnik and Stolyar [20], Stolyar [32], Stolyar [33], Stolyar and Yudovina [34], Stolyar and Yudovina [35]). It is important to understand if a queuing system is stable and the rate at which a performance measure converges to the steady state under different scheduling or routing policies. For the multiclass V-network, Gamarnik and Stolyar [20] prove the tightness of the stationary distributions of the diffusion-scaled state processes under any work-conserving scheduling policy, provided that there is $\sqrt{n}$ safety staffing (n is the scaling parameter). They show that the diffusion-scaled queuing processes are ergodic under all work-conserving scheduling policies and have exhibited exponential tail bounds for the stationary distribution. The proofs of these significant results use some natural test functions based on the total workload, but there is no uniform Foster–Lyapunov equation to exhibit the rate of convergence to the stationary distribution. For the limiting diffusion of the V-network, when the control equals $(0,\dots,0,1)^{\mathsf{T}}$, which arises as the limit of a static priority policy, the ergodic properties established in Dieker and Gao [17] for a class of piecewise Ornstein–Uhlenbeck (O-U) processes arising in many-server queues with phase-type service times can be applied. Exponential ergodicity is also established for the limiting diffusion (as a special case of a more general class of stochastic differential equations (SDEs)) under any constant control in Arapostathis et al. [9].

The following important open questions are addressed in this paper:

1. Is the limiting controlled diffusion exponentially ergodic under all stationary Markov controls? How different are the tail asymptotics of the invariant measures with or without abandonment?

2. Is there a unified approach based on Foster–Lyapunov theory that can be used to establish uniform exponential ergodicity for both the limiting diffusion and the diffusion-scaled queuing processes?

We provide affirmative answers to all these questions. We consider multiclass models with (delayed) renewal arrivals, class-dependent exponential service times, and class-dependent exponential patience times. We assume that the system is operating under work-conserving and preemptive scheduling policies. It is well known that the diffusion-scaled queuing processes under such scheduling policies converge weakly to a limiting diffusion with a drift given in Section 2.2 and a diagonal constant covariance matrix (Atar et al. [10], Harrison and Zeevi [26]).

We start with the limiting controlled diffusion. When the controls are constant, the limiting diffusion has a piecewise linear drift and belongs to a class of piecewise O-U processes. Applying Dieker and Gao [17, theorem 3] to our model with positive abandonment rate, one can deduce that the limiting diffusion is exponentially ergodic under a specific constant control corresponding to a static priority scheduling policy (Remark 1). By contrast, it is shown in Arapostathis et al. [9, theorem 3.5] that the limiting diffusion is exponentially ergodic under any constant control (Remark 2). The proofs of these results rely on the construction of a common quadratic-type Lyapunov function for the piecewise linear equations. However, this methodology only works for constant controls and leaves the question of stability over Markov controls open.

We exploit Lyapunov functions that are constructed in an intricate manner in order to capture both the total workload on the positive half-space and the idleness on the negative half-space. Such functions are, of course, quite natural and have been used as test functions in Gamarnik and Stolyar [20] to derive tail bounds. However, for the diffusion, the total workload and idleness need to be treated with the proper weights or tilting, interacting with a smoothing cutoff function that needs to be deployed. Such delicate care is not only needed for the drift as usual but, more important, also for the second-order derivatives. For multiclass queuing models in the H-W regime, such constructions appear to be necessary to deal with both the workload and idleness processes simultaneously. This constitutes our first main methodologic contribution in this paper.

We present Foster–Lyapunov equations that are uniform over all Markov controls and show that

a. If the spare capacity parameter (safety staffing) is positive, then the limiting diffusion is uniformly exponentially ergodic, and the corresponding invariant probability measures have uniform exponential tails, and

b. When the abandonment rates are all positive, regardless the spare capacity parameter being positive or negative, in addition to uniform exponential ergodicity, we show that the invariant probability measures have sub-Gaussian tails.

These answer important question 1 posed earlier.

We then show that the Foster–Lyapunov equations for the limiting diffusion offer a very natural tool with which we establish uniform ergodic properties for the diffusion-scaled queuing processes. This answers important question 2 posed earlier. In this manner, we provide a unified approach to the study of limiting diffusion and the corresponding diffusion-scaled processes.

In the case of Poisson arrivals, by employing the same Lyapunov functions used for the limiting diffusion, we show that the corresponding results in (a) hold for the diffusion-scaled queuing processes (see also Section 2.3). Sub-Gaussian tails are not possible for the invariant distribution of the diffusion-scaled queuing processes, and one can only hope for tails that decay faster than any exponential. By contrast, when the abandonment rates are all positive, we improve somewhat on the results in Gamarnik and Stolyar [20], although a conjecture stated in that paper still remains open. Although in the cases of Poisson and renewal arrivals the limiting diffusions agree, with the only differences lying in the covariance functions, for the analysis of the prelimit processes, we need to augment the state process in the renewal case.

With renewal arrivals, we consider the Markov process composed of the diffusion-scaled queuing processes and the interarrival age processes of the renewal arrivals. The Lyapunov functions used for the limiting diffusion are adapted to construct appropriate Lyapunov functions for the joint processes. By contrast, the hazard-rate functions and mean residual lifetime functions of the interarrival times also must be used in a proper manner to take into account the age processes, as suggested in Konstantopoulos and Last [28]. We prove the following results under the assumption that the residual lifetime function is bounded:

a′. If the spare capacity parameter (safety staffing) is positive, we prove a Foster–Lyapunov equation, which shows that the joint Markov process is positive Harris recurrent under any work-conserving stationary Markov scheduling policy, and

b′. If the abandonment rates are all positive, we obtain a Foster–Lyapunov equation that shows that the first absolute moments of the invariant distribution are uniformly bounded. If we impose the additional assumption that the hazard-rate function is bounded, we show that the marginal of the stationary distribution corresponding to the queuing state has exponential moments.

This work also relates to the vast literature on the validity of diffusion approximations for queues in heavy traffic. We focus on the literature of many-server queuing models in the H-W regime and refer readers to Braverman et al. [13], Budhiraja and Lee [14], Gamarnik and Zeevi [21], Gurvich [23], Katsuda [27], Ye and Yao [37], and Ye and Yao [38] and references therein for results in the conventional (single-server) heavy-traffic regime. For single-class $GI/M/n$ queues, Halfin and Whitt [25] established the interchange of limits, and they used a bounding argument via single-server queues to show the tightness of the steady-state distributions of the diffusion-scaled processes. Dai et al. [16] studied the validity of multidimensional diffusion approximations for $GI/Ph/n+M$ queues with phase-type service times. Aghajani and Ramanan [1] proved the convergence of stationary distributions of suitably scaled infinite-dimensional measure-valued processes for $GI/GI/N$ queues in the H-W regime, and they also studied the ergodic properties of the stochastic partial differential equation (SPDE) limit of the same model in Aghajani and Ramanan [2]. We also refer readers to the steady-state analysis of many-server queues in Braverman and Dai [11], Braverman et al. [12], Gamarnik and Goldberg [18], and Gamarnik and Momčilović [19]. All these studies are on single-class many-server queues. For multiclass many-server queues in the H-W regime, this topic still remains wide open. The only known result is for the Markovian N-network (Stolyar [33]), where Stolyar proves the interchange of limits for the model without abandonment under a particular static priority policy.

Uniform exponential ergodicity can substantially simplify the study of ergodic control problems because there is a rich body of existing theory that can be applied (Arapostathis et al. [7, chapter 3]). By contrast, if the system is not endowed with such blanket stability properties, and the running cost functional is not near-monotone, then the analysis of these problems can be quite involved. In the study of ergodic control of the V-network in Arapostathis et al. [6], a key structural property of the system dynamics had to be identified because of the lack of uniform stability and near-monotonicity of the running cost. It was assumed that all the abandonment rates are strictly positive, but no positive safety staffing requirement was imposed. The results in this paper enable the study of ergodic control problems for the V-network when there is no abandonment but there is positive safety staffing. Uniform stability properties are yet to be explored for multiclass multipool networks. Without such blanket stability properties, ergodic control problems for multiclass multipool networks have been recently studied by Arapostathis and Pang [3], Arapostathis and Pang [4], and Arapostathis and Pang [5] under the hypothesis that at least one abandonment rate is positive.

1.1. Notation

We summarize some of the notation used throughout this paper. We use $\mathbb{R}^m$ (and $\mathbb{R}^m_+$), $m \geq 1$, to denote real-valued m-dimensional (nonnegative) vectors and write $\mathbb{R}$ for $m = 1$. For $x, y \in \mathbb{R}$, we let $x \vee y = \max\{x, y\}$, $x \wedge y = \min\{x, y\}$, $x^+ = \max\{x, 0\}$, and $x^- = \max\{-x, 0\}$. For a set $A \subseteq \mathbb{R}^m$, we use A^c, ∂A, and 1_A to denote the complement, the boundary, and the indicator function of A, respectively. A ball of radius $r > 0$ in $\mathbb{R}^m$ around a point x is denoted by $\mathcal{B}_r(x)$ or, simply, as $\mathcal{B}_r$ if $x = 0$. We also let $\mathcal{B} \equiv \mathcal{B}_1$. The Euclidean norm on $\mathbb{R}^m$ is denoted by $|\cdot|$, and $\langle \cdot, \cdot \rangle$ stands for the inner product. Also, for $x \in \mathbb{R}^m$, we let $\|x\|_1 := \sum_i |x_i|$, $x_{\mathsf{max}} := \max_i x_i$, and $x_{\min} := \min_i x_i$, and $x^{\pm} := (x_1^{\pm}, \dots, x_m^{\pm})$. For a finite signed measure ν on $\mathbb{R}^m$ and a Borel measurable $f : \mathbb{R}^m \to [1, \infty)$, we define the f-norm of ν by

$$\|\nu\|_f := \sup_{g \in \mathcal{B}(\mathbb{R}^m),\, |g| \leq f} \left| \int_{\mathbb{R}^m} g(x)\, \nu(\mathrm{d}x) \right|, \tag{1}$$

where $\mathcal{B}(\mathbb{R}^m)$ denotes the class of Borel measurable functions on $\mathbb{R}^m$. Observe that $\|\cdot\|_1 = \|\cdot\|_{\mathsf{TV}}$, the latter denoting the total variation norm.

2. Uniform Exponential Ergodicity of the Diffusion Limit

In Section 2.1, we describe the limiting diffusion and proceed with a summary of the results and technical approach in Sections 2.2 and 2.3, respectively. Some important definitions are in Section 2.4, followed by the main technical results in Sections 2.5 and 2.6.

2.1. The Limiting Controlled Diffusion

We consider a controlled m-dimensional SDE of the form

$$\mathrm{d}X_t \;=\; b(X_t,U_t)\,\mathrm{d}t + \sigma(X(t))\,\mathrm{d}W_t\,, \qquad X(0) = x_0 \in \mathbb{R}^m, \tag{2}$$

with $b\colon\mathbb{R}^m \to \mathbb{R}^m$ given by

$$b(x,u) \;=\; \ell - M(x - \langle e,x\rangle^+ u) - \langle e,x\rangle^+ \Gamma u \;=\; \begin{cases} \ell - \big(M + (\Gamma - M)ue^{\mathsf{T}}\big)x\,, & \langle e,x\rangle > 0,\\ \ell - Mx\,, & \langle e,x\rangle \le 0. \end{cases} \tag{3}$$

Here $\ell \in \mathbb{R}^m$, $u \in \mathbb{R}^m_+$ satisfies $\langle e,u\rangle = 1$ with $e = (1,\dots,1)^{\mathsf{T}} \in \mathbb{R}^m$; $M = \mathrm{diag}(\mu_1,\dots,\mu_m)$ is a positive diagonal matrix; and $\Gamma = \mathrm{diag}(\gamma_1,\dots,\gamma_m)$ with $\gamma_i \in \mathbb{R}_+$, $i \in \mathcal{I} := \{1,\dots,m\}$. The process W_t is a standard m-dimensional Brownian motion, and the covariance function $\sigma\colon\mathbb{R}^m \to \mathbb{R}^{m\times m}$ is a positive diagonal matrix. Such a process arises as a limit of the suitably scaled queuing processes of multiclass Markovian many-server queues in the H-W regime (Atar et al. [10], Harrison and Zeevi [26]).

In these models, if the scheduling policy is based on a static priority assignment on the queues, then the vector u in (2) corresponds to a constant control that is an extreme point of the convex set

$$\Delta := \{u \in \mathbb{R}^m : u \ge 0\,,\ \langle e,u\rangle = 1\}.$$

Remark 1. As mentioned earlier, ergodicity and exponential ergodicity of a class of piecewise O-U processes as in (2) have been addressed in Dieker and Gao [17]. In this model, they assume that M is a nonsingular M-matrix such that the vector $e^{\mathsf{T}}M$ has nonnegative components, $\Gamma = \alpha I$, and $\ell = -\beta u$ for positive constants α, β, and a constant vector $u \in \Delta$. Applying their results to the multiclass $M/M/n + M$ model with abandonment, exponential ergodicity of the limiting diffusion under the specific constant control $\bar{u} = (0,\dots,0,1)^{\mathsf{T}}$, corresponding to class m being given the least priority, is established in Dieker and Gao [17, theorem 3]. By contrast, for the multiclass $M/M/n$ model without abandonment, that is, $\Gamma = 0$, positive recurrence is established for the limiting diffusion with the control $\bar{u}$, but the rate of convergence is not identified (Dieker and Gao [17, theorem 2]).

Remark 2. The model in (2) with M a nonsingular M-matrix and for constant control U_t has also been studied extensively by Arapostathis et al. [9] (as a special class of the Lévy-driven SDEs studied there). It is shown in that paper that when $\Gamma = 0$, the quantity

$$\varrho := -\,\big\langle e, M^{-1}\ell\,\big\rangle \tag{4}$$

plays a fundamental role in the characterization of stability. Specifically, it is shown in Arapostathis et al. [9, theorem 3.2] that if $\varrho > 0$, then X_t is positive recurrent under any constant control U_t, and if $\varrho < 0$ ($\varrho = 0$), then it is transient (cannot be positive recurrent) under any stationary Markov control satisfying $\Gamma v(x) = 0$ almost everywhere (a.e.) (Arapostathis et al. [9, theorem 3.3]). Another interesting property of ϱ that we find in Arapostathis et al. [9, corollary 5.1] is that provided that $\Gamma = 0$ and the diffusion under some stationary Markov control v is positive recurrent with invariant probability measure π_v, then necessarily

$$\varrho \;=\; \int_{\mathbb{R}^m} \langle e,x\rangle^- \,\pi_v(\mathrm{d}x). \tag{5}$$

This can be interpreted as follows: the average idleness in the steady state always equals the spare capacity parameter. These results, of course, apply to the problem at hand because M is a diagonal matrix. In addition, the rate of convergence is shown to be exponential if either $\Gamma u = 0$ or $\Gamma u \neq 0$ for any constant control $u \in \Delta$ (Arapostathis et al. [9, corollary 4.2]).

Let $\mathfrak{U}_{\mathsf{sm}}$ denote the class of Borel measurable maps $v\colon\mathbb{R}^m \to \Delta$. Every element v of $\mathfrak{U}_{\mathsf{sm}}$ is identified with the stationary Markov controls $U_t = v(X_t)$. Under any such control, it is well known that (2) has a unique strong solution that is a strong Feller process (Gyöngy and Krylov [24]). Let $P^v_t(x,\mathrm{d}y)$ denote its transition probability.

The diffusion in (2) is called *uniformly stable* (in the sense of Arapostathis et al. [7], definition 3.3.3]) if, under any $v \in \mathfrak{U}_{\mathsf{sm}}$, the process X_t is positive recurrent, and the collection of invariant probability measures is tight. We say that (2) is *uniformly exponentially ergodic* if it is uniformly stable and there exist positive constants C and γ and a function $\mathcal{V}\colon\mathbb{R}^m \to [1,\infty)$ such that

$$\big\|P^v_t(x,\cdot\,) - \pi_v(\cdot)\big\|_{\mathsf{TV}} \le C\mathcal{V}(x)\,\mathrm{e}^{-\gamma t}, \quad \forall (x,t) \in \mathbb{R}^m \times \mathbb{R}_+,$$

and all $v \in \mathfrak{U}_{\mathsf{sm}}$.

2.2. Brief Summary of the Results

In Theorem 1, we show that if $\varrho > 0$, then (2) is uniformly exponentially ergodic. Therefore, when $\Gamma = 0$, (5) holds over all stationary Markov controls $v \in \mathfrak{U}_{\mathsf{sm}}$. In addition, the invariant probability measures have uniform exponential tails, and by that, we mean that there exists some $\varepsilon > 0$ such that $\sup_{v\in\mathfrak{U}_{\mathsf{sm}}} \int_{\mathbb{R}^m} \mathrm{e}^{\varepsilon|x|}\,\pi_v(\mathrm{d}x) < \infty$. By contrast, if $\Gamma > 0$, then the associated invariant probability measures have sub-Gaussian tails; that is, $\sup_{v\in\mathfrak{U}_{\mathsf{sm}}} \int_{\mathbb{R}^m} \mathrm{e}^{\varepsilon|x|^2}\,\pi(\mathrm{d}x) < \infty$ for some $\varepsilon > 0$ (see Theorem 2).

In Section 3, we address the n-server networks. We first present the results for the models with (delayed) renewal arrival processes in Section 3.2. The counterpart of Theorem 1 here is given in Theorem 4, and this is established for renewal arrivals (this should be compared with Gamarnik and Stolyar [20, theorem 2]). In this theorem, the hazard-rate functions are assumed to be bounded. This is a rather strong assumption, but the result, which asserts uniform exponential tails for the invariant distributions under work-conserving stationary Markov policies, is equally strong. With strictly positive abandonment parameters, and with the hazard-rate function only locally bounded, we establish uniform stability of the queuing system under all work-conserving stationary Markov policies in Theorem 5. With possibly zero abandonment in all classes, and with positive $\sqrt{n}$ safety staffing, we show in Theorem 6 that the combined renewal age and queuing state process is positive Harris recurrent. In addition, if the limit of the safety staffing is positive, the invariant probability distributions are tight. In this result, the hazard-rate function is assumed to be only locally bounded.

Networks with Poisson arrivals are studied in Section 3.3. We show in Corollary 2 that positive spare capacity implies exponential ergodicity. However, as noted in Gamarnik and Stolyar [20], the invariant distribution of an n-server network cannot have a sub-Gaussian tail. This property is recovered only at the weak limit as $n \to \infty$, and it is worthwhile comparing Gamarnik and Stolyar [20, theorem 4] with Theorem 2, which in addition shows uniform exponential ergodicity. When all abandonment rates are positive, we can only exhibit a stronger Foster–Lyapunov equation (see Theorem 7), which implies that $\mathrm{e}^{\delta|x|}$ is uniformly integrable over the invariant probability distributions for any $\delta > 0$.

In addition to these results, we investigate the properties in Gamarnik and Stolyar [20, theorems 2(i) and 4(i)]. We provide proofs of the analogous results for the limiting diffusion in Lemma 2 and Theorem 4, respectively, using Foster–Lyapunov techniques. The counterpart of Lemma 2 for the n-system is given in Theorem 8 and is an improvement over the statement in Gamarnik and Stolyar [20, theorem 2(i)]. However, we have not been able to prove or disprove the related conjecture in Gamarnik and Stolyar [20, p. 33].

2.3. Summary of the Technical Approach

The first important step in the study of this problem is the construction of appropriate Lyapunov functions. We use two building blocks for these functions: one represents the total workload, and the other is a measure of idleness. The scaling of these in (10) plays a crucial role. Two scaling parameters are used: θ to balance the mix of workload and idleness and ε to handle the terms arising from the second derivatives in the extended generator of the controlled diffusion. Equally important are the cones in Definition 1. Although the drift of the diffusion is piecewise linear when the control is constant, it becomes quite complicated under a (stationary) Markov control. Careful analysis of the drift of the diffusion in (3) on these cones enables us to obtain the drift inequalities and Foster–Lyapunov equations in Lemma 1 and Theorems 1 and 2. The more specialized results in Section 2.6 involve Lyapunov functions that are sums of two exponentials.

The relation between the prelimit dynamics and the limiting diffusion can be described as follows. For a model with Poisson arrivals, the process $\{\hat{X}^n(t)\}_{t\geq 0}$ describing the (diffusion-scaled) total number of jobs in the system is a controlled Markov process with generator (see (68) and (93))

$$\widehat{\mathcal{A}}^n_z f(\hat{x}) := \sum_{i\in\mathcal{I}} \lambda^n_i \big(f(\hat{x} + n^{-1/2}e_i) - f(\hat{x}) \big) + \sum_{i\in\mathcal{I}} \big(\mu^n_i z_i + \gamma^n_i q_i(\hat{x}, z)\big)\big(f(\hat{x} - n^{-1/2}e_i) - f(\hat{x}) \big).$$

Here the vector $z = (z_1, \dots, z_m) \in \mathbb{Z}^m_+$ is the control parameter, with z_i denoting the number of jobs of class i in service; $\{\lambda^n_i\}_{i\in\mathcal{I}}$, $\{\mu^n_i\}_{i\in\mathcal{I}}$, and $\{\gamma^n_i\}_{i\in\mathcal{I}}$ are the arrival, service rates, and abandonment rates, respectively; and $q = (q_1, \dots, q_m)$ is the vector of queue sizes. Using the diffusion-scaled variables $\hat{z}^n$ and $\hat{q}^n$ defined in (63) as

$$\hat{z}^n_i := \frac{1}{\sqrt{n}}\left(z_i - \frac{\lambda^n_i}{\mu^n_i} \right) - \frac{\varrho^n}{m} \quad \text{and} \quad \hat{q}^n_i := \frac{q_i(x, z)}{\sqrt{n}},$$

we obtain

$$\widehat{\mathcal{A}}_z^n f(\hat{x}) = \sum_{i\in\mathcal{I}} \frac{\lambda_i^n}{n} \frac{f(\hat{x}+n^{-1/2}e_i) - 2f(\hat{x}) + f(\hat{x}-n^{-1/2}e_i)}{n^{-1}} + \sum_{i\in\mathcal{I}} \left(\mu_i^n \frac{\varrho^n}{m} + \mu_i^n \hat{z}_i^n + \gamma_i^n \hat{q}_i^n\right) \frac{f(\hat{x}-n^{-1/2}e_i) - f(\hat{x})}{n^{-1/2}}. \tag{6}$$

As shown in (88), for any work-conserving job allocation $z \in \mathbb{Z}_+^m$, there exists a vector $u \in \Delta$ such that $\hat{z}^n = \hat{x} - \langle e,\hat{x}\rangle^+ u$ and $\hat{q}^n = \langle e,\hat{x}\rangle^+ u$. Using these identities in (6) and letting $n \to \infty$, we obtain the generator of the controlled diffusion in (2) (see also Remark 3). There is some difficulty, however, with translating the Foster–Lyapunov equation for the diffusion into an analogous equation for the operator $\widehat{\mathcal{A}}_z^n$. This is because whereas $\hat{z}^n$ is of order $\sqrt{n}$, the queue sizes $\hat{q}^n$ are not bounded. We circumvent this problem by establishing drift inequalities in Lemma 1 for the truncated drift given in (15). This facilitates using the same Lyapunov function for the stability analysis of the diffusion and the prelimit, and consequently, we have a unified approach to the problem.

When studying the diffusion-scaled model with renewal arrivals, the Lyapunov function has to be augmented to account for the age processes (see (64)). The analysis is more intricate in this case, and deriving the Foster–Lyapunov equations in Theorem 4 requires extra care.

Remark 3. If we let $\zeta = \frac{\varrho}{m} e + M^{-1}\ell$, with ϱ as in (4), then a mere translation of the origin of the form $\tilde{X}_t = X_t - \zeta$ results in a diffusion of form (2) with the constant ℓ in the drift taking the form $\ell = -\frac{\varrho}{m} Me$. Therefore, without loss of generality, we assume throughout this paper that the drift in (3) takes the form

$$b(x,u) = -\frac{\varrho}{m} Me - M(x - \langle e,x\rangle^+ u) - \langle e,x\rangle^+ \Gamma u. \tag{7}$$

For $f \in \mathcal{C}^2(\mathbb{R}^m)$ and $u \in \Delta$, we define $a(x) = (a^{ij}(x))_{1\le i,j\le m} := \sigma(x)\sigma(x)^{\mathsf{T}}$ and

$$\mathcal{L}_u f(x) = \frac{1}{2}\operatorname{trace}\big(a(x)\nabla^2 f(x)\big) + \big\langle b(x,u), \nabla f(x)\big\rangle, \tag{8}$$

with $\nabla^2 f$ denoting the Hessian of f.

2.4. Preliminaries

We start with two definitions.

Definition 1. For $\delta \in [0,1]$, we define the following cones:

$$\mathcal{K}_\delta^+ := \{x \in \mathbb{R}^m : \langle e,x\rangle \ge \delta \|x\|_1\},$$
$$\mathcal{K}_\delta^- := \{x \in \mathbb{R}^m : \langle e,x\rangle \le -\delta \|x\|_1\}.$$

Note that $\mathcal{K}_0^+$ ($\mathcal{K}_0^-$) is the nonnegative (nonpositive) canonical half-space and that $\mathcal{K}_1^+$ ($\mathcal{K}_1^-$) is the nonnegative (nonpositive) closed orthant. Also, we have the following identities:

$$\langle e,x^+\rangle = \frac{1\pm\delta}{2}\|x\|_1, \qquad \langle e,x^-\rangle = \frac{1\mp\delta}{2}\|x\|_1, \qquad \text{for } x \in \partial\mathcal{K}_\delta^\pm,\ \delta \in [0,1]. \tag{9}$$

We fix some convex function $\psi \in \mathcal{C}^2(\mathbb{R})$ with the property that $\psi(t)$ is constant for $t \le -1$ and $\psi(t) = t$ for $t \ge 0$. This is defined by

$$\psi(t) := \begin{cases} -\frac{1}{2}, & t \le -1, \\ (t+1)^3 - \frac{1}{2}(t+1)^4 - \frac{1}{2}, & t \in [-1,0], \\ t, & t \ge 0. \end{cases}$$

For $\varepsilon > 0$, we define

$$\psi_\varepsilon(t) := \psi(\varepsilon t),$$

and thus $\psi_\varepsilon(t) = \varepsilon t$ for $t \ge 0$. A simple calculation also shows that $\psi_\varepsilon''(t) \le \frac{3}{2}\varepsilon^2$.

Definition 2. We let $\beta_i := \frac{\gamma_i}{\mu_i}$ for $i \in \mathcal{I}$. With θ and ε positive constants, we define

$$\Psi_\varepsilon(x) := \sum_{i\in\mathcal{I}} \frac{\psi_\varepsilon(x_i)}{\mu_i}, \quad \Psi(x) := \sum_{i\in\mathcal{I}} \frac{\psi(x_i)}{\mu_i},$$
$$\text{and} \quad \Psi^*_{\varepsilon,\theta}(x) := \varepsilon\theta\Psi(-x) + \Psi_\varepsilon(x). \tag{10}$$

The function Ψ plays a fundamental role in our analysis. The quantity $\Psi(x^+)$ represents, of course, the total workload, whereas $\Psi(x^-)$ is a measure of idleness. These functions, without the smooth cutoff part, are also used in Gamarnik and Stolyar [20] as test functions to estimate the tails of the invariant distribution of the prelimit diffusion-scaled processes.

The function $\Psi^*_{\varepsilon,\theta}$ follows the norm $\|\cdot\|_1$ in the sense that

$$\varepsilon\frac{1\wedge\theta}{\mu_{\max}}\|x\|_1 - \frac{m}{2} \le \Psi^*_{\varepsilon,\theta}(x) \le \varepsilon\frac{1\vee\theta}{\mu_{\min}}\|x\|_1. \tag{11}$$

We also have $\psi'(-1/2) = 1/2$, from which we obtain

$$\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)x_i \ge \varepsilon\|x^+\|_1 - \frac{m}{2}, \quad \text{and} \quad -\sum_{i\in\mathcal{I}}\psi'(-x_i)x_i \ge \|x^-\|_1 - \frac{m}{2}. \tag{12}$$

Note also that

$$-\varepsilon\sum_{i\in\mathcal{I}}\psi'(-x_i)x_i \le \varepsilon\langle e,x\rangle \le \sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)x_i. \tag{13}$$

Using the parameter β_i in Definition 2 and (7), we write the following identities, which we use frequently in the rest of this paper:

$$\langle\nabla\Psi_\varepsilon(x), b(x,u)\rangle = -\frac{\varrho}{m}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i) - \sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)x_i + \langle e,x\rangle^+\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)(1-\beta_i)^+u_i$$
$$- \langle e,x\rangle^+\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)(\beta_i-1)^+u_i, \tag{14a}$$
$$\langle\nabla\Psi(-x), b(x,u)\rangle = \frac{\varrho}{m}\sum_{i\in\mathcal{I}}\psi'(-x_i) + \sum_{i\in\mathcal{I}}\psi'(-x_i)x_i - \langle e,x\rangle^+\sum_{i\in\mathcal{I}}\psi'(-x_i)(1-\beta_i)^+u_i$$
$$+ \langle e,x\rangle^+\sum_{i\in\mathcal{I}}\psi'(-x_i)(\beta_i-1)^+u_i. \tag{14b}$$

2.5. Main Results on Uniform Exponential Ergodicity

The following lemma presents some important drift inequalities that are used frequently throughout this paper. Recall the definitions in (10). In order to apply the drift inequalities for the diffusion to the prelimit in Section 3, we often need to truncate the diffusion-scaled queuing processes. To prepare for this, we present a more general version of these inequalities than what is needed in this section.

For a constant $c \in [1,\infty]$, we define $b_c(x,u) := (b^1_c(x,u),\dots,b^m_c(x,u))^{\mathsf{T}}$, with

$$b^i_c(x,u) := -\frac{\varrho\mu_i}{m} - \mu_i(x_i - \langle e,x\rangle^+u_i) - \gamma_i\langle e,x\rangle^+u_i\,1_{\{x_i\le c\}}. \tag{15}$$

Lemma 1. *Assume that $\varrho > 0$, and let θ be a positive constant satisfying*

$$\theta(\beta_{\max} - 1)^+ \le 1. \tag{16}$$

Then the function

$$V(x) := \exp\bigl(\Psi^*_{\epsilon,\theta}(x)\bigr) = \exp(\varepsilon\theta\Psi(-x) + \Psi_\varepsilon(x)) \tag{17}$$

satisfies, for any constant $c \in [1,\infty]$,

$$\langle\nabla V(x), b_c(x,u)\rangle \le \begin{cases} \varepsilon\bigl(\theta\varrho + \frac{m}{2\varepsilon}(1+\varepsilon\theta) - (\theta\wedge 1)\|x\|_1\bigr)V(x), & \forall\, x \in \mathcal{K}_0^-, \\ -\varepsilon\bigl(\frac{\varrho}{m} - \theta\varrho - \theta\frac{m}{2} + \theta\|x^-\|_1\bigr)\,V(x), & \forall\,(x,u) \in \mathcal{K}_0^+ \times \Delta. \end{cases} \tag{18}$$

Proof. The bound on $\mathcal{K}_0^-$ follows by first multiplying (14b) by $\varepsilon\theta$ and then adding this equation to (14a) and using (12).

We proceed to derive the stated bound on $\mathcal{K}_0^+ \times \Delta$. Note that

$$\langle \nabla\Psi(-x), b_c(x,u)\rangle = \langle \nabla\Psi(-x), b(x,u)\rangle,$$

that is, it is equal to the right-hand side of (14b) for any $c \ge 1$, because $\psi'(-r) = 0$ for $r \ge 1$. We write

$$\langle \nabla\Psi_\varepsilon(x), b_c(x,u)\rangle = -\frac{\varrho}{m}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i) - \sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)(x_i - \langle e,x\rangle^+ u_i) - \langle e,x\rangle^+ \sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)\beta_i u_i 1_{\{x_i\le c\}}. \tag{19}$$

It holds that

$$-\frac{\varrho}{m}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i) \le -\varepsilon\frac{\varrho}{m}, \quad \text{on } \mathcal{K}_0^+. \tag{20}$$

Also, by (12), we have

$$\theta\varepsilon\frac{\varrho}{m}\sum_{i\in\mathcal{I}}\psi'(-x_i) + \theta\varepsilon\sum_{i\in\mathcal{I}}\psi'(-x_i)x_i \le \varepsilon\theta\varrho + \varepsilon\theta\frac{m}{2} - \varepsilon\theta\|x^-\|_1, \quad \text{on } \mathbb{R}^m, \tag{21}$$

and

$$\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)(x_i - \langle e,x\rangle^+ u_i) \ge 0, \quad \text{for } x\in\mathcal{K}_0^+, \tag{22}$$

by (14). Thus, if $\beta_{\max} \le 1$, then it is clear from (14a) and (19)–(22), that (18) holds for any positive ε and θ.

Next, suppose that $\beta_{\max} > 1$. We proceed by carefully comparing the terms in (19). Define

$$\widehat{\mathcal{I}} := \{i\in\mathcal{I} : \gamma_i > \mu_i\}, \quad \widehat{\mathcal{I}}_+(x) := \Big\{i\in\widehat{\mathcal{I}} : x_i \ge 0\Big\}, \quad \text{and} \quad \widehat{\mathcal{I}}_-(x) := \Big\{i\in\widehat{\mathcal{I}} : x_i < 0\Big\}. \tag{23}$$

Because $\theta(\beta_{\max}-1)^+ \le 1$ and $\beta_i > 1$ on $\widehat{\mathcal{I}}$, we have

$$\varepsilon\theta\sum_{i\in\hat{\mathcal{I}}_+(x)}\psi'(-x_i)(\beta_i-1)^+u_i \le \sum_{i\in\hat{\mathcal{I}}_+(x)}\psi'_\varepsilon(x_i)\beta_i u_i 1_{\{x_i\le c\}},$$

which we combine with

$$\sum_{i\in\hat{\mathcal{I}}_-(x)}\psi'_\varepsilon(x_i)u_i - \sum_{i\in\hat{\mathcal{I}}_-(x)}\psi'_\varepsilon(x_i)\beta_i u_i 1_{\{x_i\le c\}} \le 0,$$

to write

$$\varepsilon\theta\sum_{i\in\hat{\mathcal{I}}_+(x)}\psi'(-x_i)(\beta_i-1)^+u_i + \sum_{i\in\hat{\mathcal{I}}_-(x)}\psi'_\varepsilon(x_i)u_i - \sum_{i\in\hat{\mathcal{I}}}\psi'_\varepsilon(x_i)\beta_i u_i 1_{\{x_i\le c\}} \le 0. \tag{24}$$

By the definitions in (23), we have the identity

$$\varepsilon\sum_{i\in\hat{\mathcal{I}}_-(x)}\psi'(-x_i)u_i = \varepsilon\sum_{i\in\hat{\mathcal{I}}}u_i - \sum_{i\in\hat{\mathcal{I}}_+(x)}\psi'_\varepsilon(x_i)u_i. \tag{25}$$

Using again the fact that $\psi'(-r) = 0$ for $r \ge 1$, we obtain

$$\begin{aligned}\varepsilon\theta\langle e,x\rangle^+\sum_{i\in\hat{\mathcal{I}}_-(x)}\psi'(-x_i)(\beta_i-1)^+u_i &\le \varepsilon\theta(\beta_{\max}-1)^+\langle e,x\rangle^+\sum_{i\in\hat{\mathcal{I}}_-(x)}\psi'(-x_i)u_i\\ &\le \varepsilon\langle e,x\rangle^+\sum_{i\in\hat{\mathcal{I}}_-(x)}\psi'(-x_i)u_i\\ &\le \varepsilon\langle e,x\rangle^+\sum_{i\in\hat{\mathcal{I}}}u_i - \langle e,x\rangle^+\sum_{i\in\hat{\mathcal{I}}_+(x)}\psi'_\varepsilon(x_i)u_i\end{aligned} \tag{26}$$

for all $(x,u) \in \mathcal{K}_0^+ \times \Delta$. In the second inequality of (26), we used the fact that $\theta(\beta_{\max} - 1)^+ \le 1$, and in the third equality we used (25). Multiplying (24) by $\langle e,x\rangle^+$, adding it to (26), and then combining the resulting sum with the inequality

$$\langle e,x\rangle^+ \sum_{i\in\hat{\mathcal{I}}^c} \psi_\varepsilon'(x_i)u_i - \varepsilon\langle e,x\rangle^+ \sum_{i\in\hat{\mathcal{I}}^c} u_i \;\le\; 0,$$

where $\widehat{\mathcal{I}}^c$ denotes the complement of $\widehat{\mathcal{I}}$ with respect to $\mathcal{I}$, we obtain

$$\begin{aligned}\varepsilon\theta\langle e,x\rangle^+ &\sum_{i\in\hat{\mathcal{I}}} \psi'(-x_i)\big(\beta_i - 1\big)^+ u_i - \varepsilon\langle e,x\rangle^+ + \sum_{i\in\mathcal{I}} \psi_\varepsilon'(x_i)\langle e,x\rangle^+ u_i \\ &- \langle e,x\rangle^+ \sum_{i\in\hat{\mathcal{I}}} \beta_i\psi_\varepsilon'(x_i)u_i\, 1_{\{x_i\le c\}} \;\le\; 0.\end{aligned} \tag{27}$$

Replacing the term $-\varepsilon\langle e,x\rangle^+$ in (27) with $-\sum_{i\in\mathcal{I}}\psi_\varepsilon'(x_i)x_i$ preserves this inequality by (13). Thus, by (14b), (19)–(21), and (27), we obtain

$$\big\langle \nabla\Psi^*_{\varepsilon,\theta}(x), b_c(x,u)\big\rangle \;\le\; -\varepsilon\frac{\varrho}{m} + \varepsilon\theta\varrho + \varepsilon\theta\frac{m}{2} - \varepsilon\theta\|x^-\|_1, \qquad \forall\,(x,u)\in\mathcal{K}_0^+\times\Delta,$$

from which the second bound in (18) follows. This completes the proof. □

Recall the definitions in (1) and (8). Also recall that π_v denotes the invariant probability measure of the process governed by (2) for a control $v \in \mathfrak{U}_{\mathsf{sm}}$, under which $\{X(t)\}_{t\ge0}$ is positive recurrent.

Theorem 1. *Assume that $\varrho > 0$, and in addition to (16), let*

$$0 \;<\; \theta \;\le\; \frac{\varrho}{3m\big(2\varrho + m\big)}. \tag{28}$$

Then the following hold:

a. *There exists $\varepsilon_0 > 0$ such that for each $\varepsilon \le \varepsilon_0$, the function V in (17) satisfies the Foster–Lyapunov equation*

$$\mathcal{L}_u V(x) \;\le\; \kappa_0 - \varepsilon\Big(\frac{\varrho}{2m} + \theta\|x^-\|_1\Big)V(x),\ \forall (x,u)\in\mathbb{R}^m\times\Delta, \tag{29}$$

for some positive constant κ_0, which depends only on ε and θ. In particular, the process $\{X(t)\}_{t\ge0}$ is positive recurrent under any control $v\in\mathfrak{U}_{\mathsf{sm}}$, and

$$\int_{\mathbb{R}^m} V(x)\,\pi_v(\mathrm{d}x) \;\le\; \frac{2m}{\varepsilon\varrho}\kappa_0. \tag{30}$$

b. *There exist positive constants γ and C_γ such that*

$$\big\|P_t^v(x,\cdot\,) - \pi_v(\cdot)\big\|_V \le C_\gamma V(x)\,\mathrm{e}^{-\gamma t},\ \forall (t,x)\in\mathbb{R}_+\times\mathbb{R}^m,\ \forall\, v\in\mathfrak{U}_{\mathsf{sm}}. \tag{31}$$

Proof. Recall the definitions in (10) and also define

$$\psi^*_{\varepsilon,\theta}(t) := \varepsilon\theta\psi(-t) + \psi_\varepsilon(t), \quad t\in\mathbb{R}.$$

Write the diffusion matrix as $\sigma = \mathrm{diag}(2\tilde\lambda_1,\dots,2\tilde\lambda_m)^{1/2}$. For the queuing network, $\tilde\lambda_i = \frac{1}{2}\lambda_i(1 + c_{a,i}^2)$, where λ_i is the arrival rate (of the fluid limit), and $c_{a,i}^2$ is squared coefficient of variation of the renewal arrival process (see Section 3.1). In the case of a system with Poisson arrivals, $\tilde\lambda_i = \lambda_i$, as in (6). See Section 3.1 for the definition of these parameters. We have

$$\frac{1}{2}\,\mathrm{trace}\big(a\nabla^2 V(x)\big) \;=\; \Bigg(\sum_{i\in\mathcal{I}}\frac{\tilde\lambda_i}{\mu_i}\big(\psi^*_{\varepsilon,\theta}\big)''(x_i) + \sum_{i\in\mathcal{I}}\frac{\tilde\lambda_i}{\mu_i^2}\big[\big(\psi^*_{\varepsilon,\theta}\big)'(x_i)\big]^2\Bigg)V(x), \qquad \forall\, x\in\mathbb{R}^m.$$

Recall that $\psi_\varepsilon'' \le \frac{3}{2}\varepsilon^2$. Therefore, because also $\psi_\varepsilon' \le \varepsilon$, $\theta \le 1$, and $\sum_i \frac{\tilde\lambda_i}{\mu_i} = 1$ (see (57)), we obtain

$$\frac{1}{2}\,\mathrm{trace}\big(a\nabla^2 V(x)\big) \;\le\; \varepsilon\bigg(\frac{3}{2}(\varepsilon+\theta) + \varepsilon\overline{C}\bigg)V(x), \qquad \text{with } \overline{C} := \sum_{i\in\mathcal{I}}\frac{\tilde\lambda_i}{\mu_i^2}. \tag{32}$$

We also have $\theta\varrho + \theta\frac{m}{2} \le \frac{\varrho}{6m}$ and $\frac{3}{2}\theta \le \frac{\varrho}{4m}$ by (28). Thus, (29) follows from (18) by selecting $\varepsilon < \frac{\varrho}{6m(3+2\overline{C})}$, whereas (30) follows by (29) and Itô's formula in the usual manner.

We now turn to part (b). Write (29) as

$$\mathcal{L}_u V(x) \le \kappa_0 - \kappa_1 V(x). \tag{33}$$

We follow the proof of Meyn and Tweedie [29, theorem 6.1], which uses a δ-skeleton chain $\{X_{\delta n}\}_{n\in\mathbb{N}}$. We can use any $\delta > 0$ because $P^v_\delta(x, B) > 0$ for any set B with positive Lebesgue measure. Thus, for simplicity, we use $\delta = 1$. Then, with $t = n + s$, $s \in [0, 1)$, we have

$$\begin{aligned}\left\|P^v_t(x,\cdot) - \pi_v(\cdot)\right\|_V &= \sup_{g\in\mathcal{B}(\mathbb{R}^m),\,|g|\le V} \left|\int_{\mathbb{R}^m} P^v_{n+s}(x,\mathrm{d}y)g(y) - \int_{\mathbb{R}^m} g(y)\pi_v(\mathrm{d}y)\right| \\ &\le \int_{\mathbb{R}^m} P^v_s(x,\mathrm{d}y)\left\|P^v_n(y,\cdot) - \pi_v(\cdot)\right\|_V. \end{aligned} \tag{34}$$

Next, we estimate $\left\|P^v_n(y,\cdot) - \pi_v(\cdot)\right\|_V$ using Meyn and Tweedie [30, theorem 2.3]. Using Itô's formula and (33), we obtain (see Arapostathis et al. [7, lemma 2.5.5])

$$\int_{\mathbb{R}^m} P^v_t(x,\mathrm{d}y)V(y) = \mathbb{E}^v_x[V(X_t)] \le \frac{\kappa_0}{\kappa_1} + \mathrm{e}^{-\kappa_1 t}V(x), \qquad \forall (t,x) \in \mathbb{R}_+ \times \mathbb{R}^m,\ \forall\, v \in \mathfrak{U}_{\mathsf{sm}}. \tag{35}$$

Therefore, with $\mathcal{B}$ a ball such that $\frac{\kappa_0}{\kappa_1} \le \mathrm{e}^{-\frac{\kappa_1}{2}}(1 - \mathrm{e}^{-\frac{\kappa_1}{2}})V(x)$ for $x \in \mathcal{B}^c$, we have

$$\int_{\mathbb{R}^m} P^v_1(x,\mathrm{d}y)V(y) \le \mathrm{e}^{-\frac{\kappa_1}{2}}V(x) + \frac{\kappa_0}{\kappa_1}1_{\mathcal{B}}(x),$$

which establishes equation (14) in Meyn and Tweedie [30].

The inequality in (30) implies that the collection of invariant probability measures $\{\pi_v : v \in \mathfrak{U}_{\mathsf{sm}}\}$ is tight. By the invariance of π_v, tightness, and the Harnack inequality applied to the densities of π_v (see Arapostathis et al. [7, lemma 3.2.4(b)]), we have

$$\int_{\mathbb{R}^m} \pi_v(\mathrm{d}y)P^v_{1/2}(y,\mathcal{B}) = \pi_v(\mathcal{B}) \ge \beta_0 > 0$$

for some constant β_0 independent of $v \in \mathfrak{U}_{\mathsf{sm}}$. Using tightness once more, we can select a ball $B_R \supset \mathcal{B}$ such that

$$\int_{B_R} \pi_v(\mathrm{d}y)P^v_{1/2}(y,\mathcal{B}) \ge \frac{\beta_0}{2}.$$

This implies that $\sup_{y\in B_R} P^v_{1/2}(y,\mathcal{B}) \ge \frac{\beta_0}{2}$. We now use the parabolic Harnack inequality for operators in nondivergence form (Gruber [22, theorem 4.1]; for a simpler statement that uses the notation in this paper, see Arapostathis et al. [8, theorem 4.7]). The parabolic Harnack inequality asserts that there exists a positive constant C_{H} such that

$$\sup_{y\in B_R} P^v_{1/2}(y,\mathcal{B}) \le C_{\mathsf{H}} \inf_{y\in B_R} P^v_1(y,\mathcal{B}), \qquad \forall\, v \in \mathfrak{U}_{\mathsf{sm}}.$$

Therefore, $P^v_1(x,\mathcal{B}) \ge \frac{1}{2}C_{\mathsf{H}}^{-1}\beta_0$ for all $x \in \mathcal{B}$ and $v \in \mathfrak{U}_{\mathsf{sm}}$. Thus, with $\delta_0 := \frac{1}{4}C_{\mathsf{H}}^{-1}\beta_0$, we can write

$$\eta := \inf_{y\in\mathcal{B}} P^v_1(y,\mathcal{B}) - \delta_0 \ge \delta_0 \qquad \forall\, v \in \mathfrak{U}_{\mathsf{sm}},$$

which establishes equation (23) in Meyn and Tweedie [30].

As seen then in Meyn and Tweedie [30, theorem 2.3, equations (19), (20), and (24)–(25)], there exist positive constants C_0 and γ depending only on κ_0, κ_1, η, and δ_0 such that

$$\left\|P^v_n(x,\cdot) - \pi_v(\cdot)\right\|_V \le C_0\,\mathrm{e}^{-\gamma n}V(x). \tag{36}$$

Thus, using (36) in (34) and applying (35) once more, we obtain (31) for a constant C_γ independent of $v \in \mathfrak{U}_{\mathsf{sm}}$. This completes the proof. □

Throughout this paper, we let K_r, or $K(r)$, for $r > 0$, denote the closed cube

$$K_r := \{x \in \mathbb{R}^m : \|x\|_1 \le r\}. \tag{37}$$

We also let $\overline{\psi}_\varepsilon = \psi_\varepsilon + 1$ so that the function is strictly positive and define $\overline{\Psi}$ and $\overline{\Psi}_\varepsilon$ analogously to (10).

Remark 4. Assume that $\varrho > 0$, and consider the function

$$\mathcal{V}(x) := \big(\varepsilon\theta\overline{\Psi}(-x) + \overline{\Psi}_\varepsilon(x)\big)^p \tag{38}$$

for some $p \ge 1$. Then it follows directly from the proofs of Lemma 1 and Theorem 1 that there exist positive constants ε, θ, $\overline{\kappa}_0$, and $\overline{\kappa}_1$ and a cube $K \subset \mathbb{R}^m$ depending only on p such that

$$\mathcal{L}_u\mathcal{V}(x) \le \begin{cases} \overline{\kappa}_0\,1_K(x) - \overline{\kappa}_1\mathcal{V}(x), & \forall\, x \in \mathcal{K}_0^-, \\ -p\varepsilon\frac{\varrho}{2m}\,(\mathcal{V}(x))^{\frac{p-1}{p}}, & \forall\,(x,u) \in \mathcal{K}_0^+ \times \Delta. \end{cases}$$

In Theorem 2, we do not assume that $\varrho > 0$.

Theorem 2. *Assume that $\Gamma > 0$. With $\overline{C}$ as defined in (32), let*

$$\theta = \frac{(1-\beta_{\min}) \vee \frac{1}{2}}{\beta_{\max}} \quad \text{and} \quad \varepsilon_0 := \frac{1}{2\sqrt{\overline{C}}}\left[\theta \wedge \beta_{\min}\left(\beta_{\min} \wedge \frac{1}{2}\right)\right]\frac{(1\wedge\theta)\mu_{\min}}{(1\vee\theta)^2\mu_{\max}}. \tag{39}$$

Then, for any $\varepsilon \le \varepsilon_0$, the function

$$\widetilde{V}(x) := \exp\left(\frac{1}{2}\left[\Psi^*_{\varepsilon,\theta}(x)\right]^2\right) = \exp\left(\frac{1}{2}\left[\varepsilon\theta\Psi(-x) + \Psi_\varepsilon(x)\right]^2\right)$$

satisfies the Foster–Lyapunov equation

$$\mathcal{L}_u\widetilde{V}(x) \le \tilde{\kappa}_0 - \varepsilon^2\left[\theta \wedge \beta_{\min}\left(\beta_{\min} \wedge \frac{1}{2}\right)\right]\frac{1\wedge\theta}{2\mu_{\max}}\|x^2\|_1\,\widetilde{V}(x),\ \forall (x,u) \in \mathbb{R}^m \times \Delta, \tag{40}$$

for a positive constant $\tilde{\kappa}_0$ that depends only on ε and the system parameters. In particular, the process X_t governed by (2) is uniformly exponentially ergodic, and the associated invariant probability measures have sub-Gaussian tails.

Proof. We borrow some calculations from the proof of Lemma 1. Using (25) and scaling this with the new definition of θ in (39), we have

$$\begin{aligned}\left((1-\beta_{\min}) \vee \frac{1}{2}\right)&\left(-\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)x_i + \langle e,x\rangle\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)(1-\beta_i)^+u_i\right) \\ &+ \varepsilon\theta\langle e,x\rangle\sum_{i\in\mathcal{I}}\psi'(-x_i)(\beta_i-1)^+u_i \le 0.\end{aligned} \tag{41}$$

Here ϱ is not necessarily positive, so by (12), we have

$$-\frac{\varrho}{m}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i) + \varepsilon\theta\frac{\varrho}{m}\sum_{i\in\mathcal{I}}\psi'(-x_i) + \varepsilon\theta\sum_{i\in\mathcal{I}}\psi'(-x_i)x_i \le \varepsilon\left(|\varrho| + \theta|\varrho| + \theta\frac{m}{2} - \theta\|x^-\|_1\right) \tag{42}$$

on $\mathbb{R}^m$. Note that

$$\langle e,x\rangle\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)(1-\beta_i)^+u_i \le \|x^+\|_1(1-\beta_{\min})\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)u_i \le \|x^+\|_1(1-\beta_{\min}).$$

Thus, using (12), we have

$$\left(\beta_{\min} \wedge \frac{1}{2}\right)\left(-\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)x_i + \langle e,x\rangle\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)(1-\beta_i)^+u_i\right) \le \varepsilon\left(\beta_{\min} \wedge \frac{1}{2}\right)\left(\frac{m}{2\varepsilon} - \beta_{\min}\|x^+\|_1\right). \tag{43}$$

Let $\overline{\theta}:=\theta\wedge\beta_{\min}(\beta_{\min}\wedge\frac{1}{2})$. Adding (41)–(43), using (14a) and (14b) and also (11), we obtain

$$\begin{aligned}
\Psi^*_{\varepsilon,\theta}(x)\big\langle\nabla\Psi^*_{\varepsilon,\theta}(x),b(x,u)\big\rangle &\le \varepsilon\left(|\varrho|+\theta|\varrho|+\theta\frac{m}{2}+\left(\beta_{\min}\wedge\frac{1}{2}\right)\frac{m}{2\varepsilon}-\overline{\theta}\|x\|_1\right)\Big)\Psi^*_{\varepsilon,\theta}(x)\\
&\le \varepsilon\left(|\varrho|+\theta|\varrho|+\theta\frac{m}{2}+\left(\beta_{\min}\wedge\frac{1}{2}\right)\frac{m}{2\varepsilon}\right)\frac{1\vee\theta}{\mu_{\min}}\|x\|_1\\
&\quad-\varepsilon^2\overline{\theta}\left(\frac{1\wedge\theta}{\mu_{\max}}\|x\|_1-\frac{m}{2\varepsilon}\right)\|x\|_1\\
&\le \varepsilon\hat{c}_0-\varepsilon^2\overline{\theta}\frac{1\wedge\theta}{\mu_{\max}}\|x\|_1^2,\qquad\forall\,(x,u)\in\mathcal{K}_0^+\times\Delta,
\end{aligned}\tag{44}$$

where

$$\hat{c}_0:=\left(|\varrho|+\theta|\varrho|+\theta\frac{m}{2}+\left(\beta_{\min}\wedge\frac{1}{2}\right)\frac{m}{2\varepsilon}\right)\frac{1\vee\theta}{\mu_{\min}}+\frac{m}{2}\bar{\theta}.$$

It is straightforward to verify that (44) is also valid on $\mathcal{K}_0^-\times\Delta$. Following the proof of Theorem 1, we have

$$\begin{aligned}
\operatorname{trace}\left(a\nabla^2\widetilde{V}(x)\right) &\le \left[\frac{3}{2}\varepsilon(\varepsilon+\theta)\Psi^*_{\varepsilon,\theta}(x)+\varepsilon^2(1\vee\theta)^2\overline{C}\left(1+(\Psi^*_{\varepsilon,\theta}(x))^2\right)\right]\widetilde{V}(x)\\
&\le \varepsilon\left[\varepsilon(1\vee\theta)^2\overline{C}+\frac{3}{2}\varepsilon(\varepsilon+\theta)\frac{1\vee\theta}{\mu_{\min}}\|x\|_1+\varepsilon^3\overline{C}\frac{(1\vee\theta)^4}{\mu_{\min}^2}\|x\|_1^2\right]\widetilde{V}(x).
\end{aligned}\tag{45}$$

Combining (44) and (45), we obtain

$$\begin{aligned}
\operatorname{trace}\left(a\nabla^2\widetilde{V}(x)\right)+\left\langle\nabla\widetilde{V}(x),b(x,u)\right\rangle &\le \Bigg[\varepsilon^2(1\vee\theta)^2\overline{C}+\varepsilon\left(\frac{3}{2}(\varepsilon+\theta)\frac{1\vee\theta}{\mu_{\min}}+\hat{c}_0\right)\|x\|_1\\
&\quad-\varepsilon^2\left(\bar{\theta}\frac{1\wedge\theta}{\mu_{\max}}-\varepsilon^2\overline{C}\frac{(1\vee\theta)^4}{\mu_{\min}^2}\right)\|x^2\|_1\Bigg]\widetilde{V}(x),
\end{aligned}$$

from which the validity of (40) on $\mathcal{K}_0^+\times\Delta$ follows by selecting ε sufficiently small. Verifying the validity of (40) on $\mathcal{K}_0^-\times\Delta$ is simpler and is a straightforward application of (11), (12), and (45). This finishes the proof. □

Remark 5. The counterpart of Remark 4 applies relative to Theorem 2. In particular, the function $\mathcal{V}$ in (38) for $p>0$ is a Lyapunov function. Indeed, there exist positive constants ε, θ, $\hat{\kappa}_0$, and $\hat{\kappa}_1$ and a cube $K\subset\mathbb{R}^m$ depending only on p such that

$$\mathcal{L}_u\mathcal{V}(x)\le\hat{\kappa}_0\,1_K(x)-\hat{\kappa}_1\mathcal{V}(x)\qquad,\forall\,(x,u)\in\mathcal{K}_0^+\times\Delta.$$

Remark 6. It is worth noting that if $\Gamma>0$, then, by choosing $\theta>0$ as in (39), the function

$$\check{V}(x):=\exp\big(\eta\theta\Psi(-x)+\eta\Psi(x)\big)$$

satisfies

$$\mathcal{L}_u\check{V}(x)\le\check{\kappa}_0-\check{\kappa}_1\|x\|_1\check{V}(x)\qquad,\forall\,(x,u)\in\mathcal{K}_0^+\times\Delta,$$

for all $\eta>0$ and for some positive constants $\check{\kappa}_0$ and $\check{\kappa}_1$ depending only on η. Indeed, using (14a), (14b), and (41), we deduce, with $\hat{\theta}:=1-((1-\beta_{\min})\vee\frac{1}{2})$, that

$$\begin{aligned}
\frac{1}{\eta\check{V}(x)}\langle\nabla\check{V}(x),b(x,u)\rangle &= \frac{\varrho}{m}\sum_{i\in\mathcal{I}}\big(\theta\psi'(-x_i)-\psi'(x_i)\big)-\sum_{i\in\mathcal{I}}\big((1-\hat{\theta})\psi'_\varepsilon(x_i)x_i-\theta\psi'_\varepsilon(-x_i)\big)x_i\\
&\quad+\big(1-\hat{\theta}\big)\langle e,x\rangle^+\sum_{i\in\mathcal{I}}\psi'(x_i)\big(1-\beta_{\min}\big)^+u_i\\
&\le\frac{m}{2}\big(1-\hat{\theta}+\theta\big)+\frac{\varrho}{m}\sum_{i\in\mathcal{I}}\big(\theta\psi'(-x_i)-\psi'(x_i)\big)-\big(\beta_{\min}\big(1-\hat{\theta}\big)\wedge\theta\big)\|x\|_1,
\end{aligned}$$

where we also used (12) and (13). The rest is routine.

2.6. Results Concerning the Tail of the Invariant Distribution

Gamarnik and Stolyar [20] conjecture that provided that $\varrho > 0$, $\exp(\theta \sum_i x_i^-)$ is integrable under an invariant probability measure for all $\theta > 0$. They prove this when $\gamma_i \le \mu_i$ for all $i \in \mathcal{I}$ (Gamarnik and Stolyar [20, theorem 2(i)]. The proof is for the diffusion-scaled queuing processes and relies on a simple comparison with a system with infinitely many servers. For this proof to go through, however, it seems necessary that all i satisfy $\gamma_i \le \mu_i$. We improve on this result by showing that $\mathrm{e}^{\theta x_i^-}$ is integrable under an invariant probability measure for all $\theta > 0$ for any i such that $\gamma_i \le \mu_i$. Of course, this proof applies to the limiting diffusion, but we show in Section 3 how to recover this property for the prelimit in Theorem 8. The general conjecture remains open.

We need some notation. We let

$$\mathcal{I}_1 := \bigl\{i \in \mathcal{I} : \gamma_i \le \mu_i\bigr\}, \tag{46}$$

and for a positive constant η, we define

$$\Phi_1(x) := \sum_{i \in \mathcal{I}_1} \frac{\psi(-x_i)}{\mu_i} \quad \text{and} \quad \mathcal{V}_1(x) := \exp\bigl(\eta \Phi_1(x)\bigr). \tag{47}$$

Lemma 2. *Assume that $\varrho > 0$. Let $\eta > 0$ be arbitrary, and $V(x) = \exp(\Psi^*_{\varepsilon_0,\theta}(x))$, with ε_0 as in Theorem 1, and the constant θ chosen to satisfy (16) and (28). Then*

$$\mathcal{L}_u(\mathcal{V}_1 + V)(x) \le \begin{cases} \kappa_0\, 1_K(x) - \kappa_1 \|x\|_1 (\mathcal{V}_1(x) + V(x)), & \forall\, x \in \mathcal{K}_0^-, \\ \kappa_0\, 1_K(x) - \varepsilon_0 \frac{\varrho}{8m} (\mathcal{V}_1(x) + V(x)), & \forall\, (x,u) \in \mathcal{K}_0^+ \times \Delta \end{cases}$$

for some positive constants κ_0 and κ_1 and some cube $K \in \mathbb{R}^m$.

Proof. Using (14a) and (14b), we write

$$\begin{aligned} \frac{1}{\mathcal{V}_1(x)} \langle \nabla \mathcal{V}_1(x), b(x,u) \rangle &= \frac{1}{2}\eta |\mathcal{I}_1| + \eta \frac{\varrho}{m} \sum_{i \in \mathcal{I}_1} \psi'(-x_i) - \eta \biggl(\frac{1}{2}|\mathcal{I}_1| - \sum_{i \in \mathcal{I}_1} \psi'(-x_i) x_i\biggr) \\ &\quad - \eta \langle e, x\rangle^+ \sum_{i \in \mathcal{I}_1} (1 - \beta_i)\, \psi'(-x_i) u_i. \end{aligned} \tag{48}$$

Let

$$H(x) := \operatorname{trace}\bigl(a \nabla^2 \Phi(x)\bigr) + \bigl\langle \nabla\Phi(x), a\bigl(\nabla\Phi(x) + 2\nabla\Psi^*_{\varepsilon,\theta}(x)\bigr)\bigr\rangle.$$

Recall the definition in (37). It is clear from (9) that we can select $\delta \in (0,1)$ and $r > 0$ such that

$$\begin{aligned} H(x)\mathcal{V}_1(x) + \Bigl(\eta \frac{m}{2} + \eta\varrho\Bigr) \mathcal{V}_1(x) &\le \varepsilon_0 \frac{\varrho}{4m} V(x) \\ \text{and } \ \mathcal{V}_1(x) &\le V(x), \qquad \forall\, x \in K_r^{\mathrm{c}} \cap \mathcal{K}_\delta^+. \end{aligned} \tag{49}$$

Combining (29) and (49), we obtain

$$\mathcal{L}_u(\mathcal{V}_1 + V)(x) \le \kappa_0 - \frac{\varepsilon_0}{2}\Bigl(\frac{\varrho}{4m} + \theta \|x^-\|_1\Bigr)(\mathcal{V}_1(x) + V(x)), \qquad \forall\, x \in K_r^{\mathrm{c}} \cap \mathcal{K}_\delta^+, \tag{50}$$

and all $u \in \Delta$. By (48), we have

$$\frac{1}{\mathcal{V}_1(x)} \langle \nabla \mathcal{V}_1(x), b(x,u)\rangle \le \eta\Bigl(\frac{m}{2} + \varrho\Bigr) - \eta \sum_{i \in \mathcal{I}_1} x_i^-. \tag{51}$$

Consider the set

$$\mathcal{K} := \biggl\{x \in \mathcal{K}_0^+ \setminus \mathcal{K}_\delta^+ : \frac{1}{2}\eta \sum_{i \in \mathcal{I}_1} x_i^- \le \eta\Bigl(\frac{m}{2} + \varrho\Bigr) + H(x) + \varepsilon_0 \frac{\varrho}{4m}\biggr\}.$$

Because H is bounded on $\mathbb{R}^m$, it is clear by the definition of $\mathcal{K}$ that $\mathcal{V}_1$ and $\mathcal{L}_u \mathcal{V}_1$ are both bounded on $\mathcal{K}$. Therefore, because V is coercive on $\mathcal{K}$ (i.e., $\liminf_{\{|x| \to \infty,\, x \in \mathcal{K}\}} V(x) \to \infty$), there exists $r_\circ > 0$ such that

$$|\mathcal{L}_u \mathcal{V}_1(x)| \le \varepsilon_0 \frac{\varrho}{4m} V(x) \quad \text{and} \quad \mathcal{V}_1(x) \le \mathcal{V}(x), \qquad \forall (x,u) \in \Bigl(\mathcal{K} \cap K_{r_\circ}^{\mathrm{c}}\Bigr) \times \Delta. \tag{52}$$

By contrast, we have

$$\mathcal{L}_u\mathcal{V}_1(x) \le -\Bigl(\varepsilon_0\,\frac{\varrho}{4m} + \frac{\eta}{2}\|x^-\|_1\Bigr)\mathcal{V}_1(x), \quad \forall\,(x,u) \in \bigl(\mathcal{K}_0^+ \setminus \mathcal{K}_\delta^+\bigr) \cap \mathfrak{K}^{c} \tag{53}$$

by (51). Equations (52) and (53), together with (29) and (50), imply that

$$\mathcal{L}_u(\mathcal{V}_1 + V)(x) \le \kappa_0 - \frac{\varepsilon_0}{2}\Bigl(\frac{\varrho}{4m} + \Bigl(\theta \wedge \frac{\eta}{2}\Bigr)\|x^-\|_1\Bigr)(\mathcal{V}_1(x) + V(x)), \qquad \forall\, x \in K^{c}_{r\vee r_\circ} \cap \mathcal{K}_0^+.$$

The estimate on $\mathcal{K}_0^-$ is straightforward. Indeed, (48) shows that $\mathcal{V}_1$ satisfies this estimate, and (29) asserts the same for V. This completes the proof. □

The following is immediate from Lemma 2.

Corollary 1. *Suppose that $\varrho > 0$. Then the function $\exp(\eta \sum_{i\in\mathcal{I}_1} \frac{\psi(-x_i)}{\mu_i})$ is integrable under the invariant distribution for any $\eta > 0$.*

In Gamarnik and Stolyar [20, theorem 4(i)], it is shown that if ν is any limit of the invariant distributions of the diffusion-scaled queuing processes, then there exists some θ such that $f(x) = \exp(\theta \sum_i (x_i^-)^2)$ is integrable under ν. As is pointed out in Gamarnik and Stolyar [20], this property holds only at the limit. The function f is not integrable under the stationary distribution of the prelimit model. The proof is rather tedious and is approached via truncations (see Gamarnik and Stolyar [20, proposition 12]). In what follows, we provide a simple proof of this result by showing that this property holds for the limiting diffusion.

Recall the definitions in (46) and (47).

Theorem 3. *Assume that $\varrho > 0$, and let*

$$\Phi_\eta(x) := \sum_{i\in\mathcal{I}_1} \frac{\psi_\eta(-x_i)}{\mu_i}, \qquad \widetilde{\mathcal{V}}_\eta(x) := \exp\Bigl(\frac{1}{2}\bigl[\eta\Phi_\eta(x)\bigr]^2\Bigr), \quad \text{and} \quad V(x) := \exp\Bigl(\Psi^*_{\varepsilon_0,\theta}(x)\Bigr),$$

with ε_0 and θ chosen as in Lemma 2. *Then there exists $\eta > 0$ such that the function $\mathcal{V} := \widetilde{\mathcal{V}}_\eta V$ satisfies*

$$\mathcal{L}_u\mathcal{V}(x) \le c_0 - c_1\mathcal{V}(x), \qquad \forall\,(x,u) \in \mathbb{R}^m \times \mathbb{U}.$$

Proof. As in (48), we have

$$\begin{aligned}\langle \nabla\Phi_\eta(x), b(x,u)\rangle &= \frac{1}{2}|\mathcal{I}_1| + \frac{\varrho}{m}\sum_{i\in\mathcal{I}_1}\psi'_\eta(-x_i) - \Bigl(\frac{1}{2}|\mathcal{I}_1| - \sum_{i\in\mathcal{I}_1}\psi'_\eta(-x_i)x_i\Bigr)\\ &\quad - \langle e, x\rangle^+ \sum_{i\in\mathcal{I}_1}(1-\beta_i)\,\psi'_\eta(-x_i)u_i, \qquad \forall\,(x,u) \in \mathbb{R}^m \times \Delta.\end{aligned} \tag{54}$$

Let

$$\widetilde{H}_\eta(x) := \frac{1}{2}\operatorname{trace}\Bigl(a\nabla^2[\Phi_\eta(x)]^2\Bigr) + \frac{1}{2}\Bigl\langle \nabla[\Phi_\eta(x)]^2, a\Bigl(\nabla[\Phi_\eta(x)]^2 + 2\nabla\Psi^*_{\varepsilon,\theta}(x)\Bigr)\Bigr\rangle.$$

Note that $\widetilde{H}_\eta(x) \le c_0\eta^2 + c_1\eta^4[\Phi_\eta(x)]^2$ for some positive constants $c_0 + c_1$. Consider the set

$$\widetilde{\mathfrak{K}} := \Bigl\{x \in \mathcal{K}_0^+ : \eta\sum_{i\in\mathcal{I}_1} x_i^- \le \frac{|\mathcal{I}_1|}{2} + \eta\varrho + \Bigl(\eta^2\widetilde{H}_\eta(x) + \varepsilon_0\,\eta\frac{\varrho}{4m}\Bigr)[\Phi_\eta(x)]^{-1}\Bigr\}.$$

It is clear that $\Phi_\eta(x)$ is bounded on this set, and thus the same applies to $\widetilde{H}_\eta$ and $\langle\nabla\Phi_\eta(x), b(x,u)\rangle$. Thus, we have

$$\sup_{x\in\widetilde{\mathfrak{K}}}\Bigl[\eta^2\widetilde{H}_\eta(x) + \eta^2\Phi_\eta(x)\bigl\langle\nabla\Phi_\eta(x), b(x,u)\bigr\rangle\Bigr] \xrightarrow[\eta\searrow 0]{} 0. \tag{55}$$

However, (29) and (55) imply that η may be selected small enough that

$$\mathcal{L}_u\mathcal{V}(x) \le \kappa_0 - \varepsilon\Bigl(\frac{\varrho}{4m} + \theta\|x^-\|_1\Bigr)\mathcal{V}(x), \qquad \forall (x,u) \in \Bigl(\mathcal{K}_0^+ \cap \widetilde{\mathfrak{K}}\Bigr) \times \Delta. \tag{56}$$

By contrast, by (54) and the definition of $\widetilde{\mathcal{K}}$, we have

$$\eta^2 \widetilde{H}_\eta(x) + \eta^2 \Phi_\eta(x) \langle \nabla \Phi_\eta(x), b(x,u) \rangle \le 0, \qquad \forall\, (x,u) \in \left(\mathcal{K}_0^+ \cap \widetilde{\mathcal{K}}^c\right) \times \Delta,$$

which also implies (56) on $(\mathcal{K}_0^+ \cap \widetilde{\mathcal{K}}^c) \times \Delta$. Because the bound on $\mathcal{K}^+$ is clear, this completes the proof. □

3. Uniform Ergodicity of Multiclass Many-Server Queues

For a detailed description of this model, see Arapostathis et al. [3]. Here we only review the basic structure that is used for our results. We consider a sequence of $GI/M/n+M$ queues with m classes of customers, indexed by n, which is the number of servers. Customers of each class form their own queue and are served in the order of their arrival.

3.1. Model and Assumptions

Let A_i^n, $i \in \mathcal{I} = \{1, \ldots, m\}$, denote the arrival process of class-i customers with arrival rate λ_i^n. We assume that $\{A_i^n\}_{i\in\mathcal{I}}$ are renewal processes defined as follows. Let $\{R_{ij} : i \in \mathcal{I}, j \in \mathbb{N}\}$ be a collection of independent positive random variables such that for each $i \in \mathcal{I}$, $\{R_{ij}\}_{j\in\mathbb{N}}$ have a common distribution function F_i having a density f_i, mean equal to one, and squared coefficient of variation (SCV) $c_{a,i}^2 \in (0,\infty)$. Let

$$h_i(\tau) := \frac{f_i(\tau)}{1 - F_i(\tau)} \quad \text{and} \quad \zeta_i(\tau) := \frac{\int_\tau^\infty (1 - F_i(r))\,\mathrm{d}r}{1 - F_i(\tau)}$$

for $\tau \ge 0$, and denote the hazard rate and the mean residual life functions for each $i \in \mathcal{I}$, respectively. The arrival process A_i^n is then given by

$$A_i^n(t) := \max\left\{k \ge 0 : \sum_{j=1}^k R_{ij} \le \lambda_i^n t\right\}, \quad t \ge 0,\, i \in \mathcal{I}.$$

We assume the following structural hypotheses on the collection $\{F_i\}_{i\in\mathcal{I}}$, which are enforced in this subsection without further mention.

Assumption 1. *The distribution functions $\{F_i\}_{i\in\mathcal{I}}$ satisfy $F_i(0) = 0$ and have a locally bounded density f_i with unbounded support. In addition, the mean residual life functions $\{\zeta_i\}_{i\in\mathcal{I}}$ are bounded.*

The service and patience times are exponentially distributed, with class-dependent rates μ_i and γ_i, respectively, for class-i customers. The arrival, service, and abandonment processes of each class are mutually independent.

The queuing process (counting the number both in service and in queue for each class) of the nth system $X^n = \{X^n(t) : t \ge 0\}$ is governed by

$$X_i^n(t) = X_i^n(0) + A_i^n(t) - Y_i^n\left(\mu_i^n \int_0^t Z_i^n(s)\mathrm{d}s\right) - R_i^n\left(\gamma_i^n \int_0^t Q_i^n(s)\mathrm{d}s\right)$$

for $i \in \mathcal{I}$ and $t \ge 0$. Here Y_i^n and R_i^n, are mutually independent rate-1 Poisson processes, independent of the initial conditions $X_i^n(0)$ and the arrival processes A_i^n, for all $i \in \mathcal{I}$. Also, $Z_i^n(s)$ and $Q_i^n(s)$ represent the numbers of class-i jobs in service and in queue at time s, $s \ge 0$, respectively.

3.1.1. The H-W regime. The parameters satisfy the following limits as $n \to \infty$ for all $i \in \mathcal{I}$:

$$\frac{\lambda_i^n}{n} \to \lambda_i > 0, \qquad \mu_i^n \to \mu_i > 0, \qquad \gamma_i^n \to \gamma_i \ge 0,$$
$$\frac{\lambda_i^n - n\lambda_i}{\sqrt{n}} \to \hat{\lambda}_i, \qquad \sqrt{n}\,(\mu_i^n - \mu_i) \to \hat{\mu}_i,$$
$$\rho_i^n := \frac{\lambda_i^n}{n\mu_i^n} \to \rho_i := \frac{\lambda_i}{\mu_i} < 1, \qquad \sum_{i=1}^m \rho_i = 1. \tag{57}$$

The assumptions in (57) imply that

$$\varrho^n := \sqrt{n}\left(1 - \sum_{i=1}^{m} \frac{\lambda_i^n}{n\mu_i^n}\right) \to \varrho := \sum_{i=1}^{m} \frac{\rho_i \hat{\mu}_i - \hat{\lambda}_i}{\mu_i} \in \mathbb{R}. \tag{58}$$

We define the diffusion-scaled variables by

$$\begin{aligned} \hat{X}_i^n(t) &= \frac{1}{\sqrt{n}}\left(X_i^n(t) - \frac{\lambda_i^n}{\mu_i^n}\right) - \frac{\varrho^n}{m}, \quad \hat{Z}_i^n(t) = \frac{1}{\sqrt{n}}\left(Z_i^n(t) - \frac{\lambda_i^n}{\mu_i^n}\right) - \frac{\varrho^n}{m}, \\ \hat{Q}_i^n(t) &= \frac{1}{\sqrt{n}} Q_i^n(t), \quad \text{and} \quad \hat{A}_i^n(t) = \frac{1}{\sqrt{n}}\big(A_i^n(t) - \lambda_i^n t\big), \quad i \in \mathcal{I}. \end{aligned} \tag{59}$$

Then we obtain the following representation of $\hat{X}_i^n(t)$:

$$\begin{aligned} \hat{X}_i^n(t) &= \hat{X}_i^n(0) - \frac{\varrho^n \mu_i^n}{m} t - \mu_i^n \int_0^t \hat{Z}_i^n(s)\,\mathrm{d}s - \gamma_i^n \int_0^t \hat{Q}_i^n(s)\,\mathrm{d}s \\ &\quad + \hat{A}_i^n(t) - \hat{M}_{S,i}^n(t) - \hat{M}_{R,i}^n(t), \quad t \geq 0, \end{aligned} \tag{60}$$

where

$$\begin{aligned} \hat{M}_{Y,i}^n(t) &:= \frac{1}{\sqrt{n}}\left(Y_i^n\left(\mu_i^n \int_0^t Z_i^n(s)\mathrm{d}s\right) - \mu_i^n \int_0^t \hat{Z}_i^n(s)\,\mathrm{d}s\right), \\ \hat{M}_{R,i}^n(t) &:= \frac{1}{\sqrt{n}}\left(R_i^n\left(\gamma_i^n \int_0^t Q_i^n(s)\mathrm{d}s\right) - \gamma_i^n \int_0^t Q_i^n(s)\mathrm{d}s\right), \end{aligned}$$

and the last two terms $\hat{M}_{Y,i}^n(t)$ and $\hat{M}_{R,i}^n(t)$ are square integrable martingales associated with the service and abandonment processes, respectively. The martingales are compensated rate-1 Poisson processes with random time changes with respect to the natural filtration (Arapostathis et al. [6]).

The diffusion-scaled arrival processes satisfy

$$\hat{A}^n \Rightarrow \operatorname{diag}\big(\lambda_1 c_{a,1}^2, \ldots, \lambda_m c_{a,m}^2\big)^{1/2} W \quad \text{in } (\mathbb{D}_m, J_1) \quad \text{as } n \to \infty,$$

where W is a standard m-dimensional Wiener process, and $(\mathbb{D}_m, J_1)$ represents the space of càdlàg functions in $\mathbb{R}^m$ endowed with the Skorokhod J_1 topology. Assuming that $\hat{X}^n(0) \Rightarrow X(0) = x_0$ for a constant $x_0 \in \mathbb{R}^m$, it then follows that $\hat{X}^n \Rightarrow X$ in $(\mathbb{D}_m, J_1)$ as $n \to \infty$, where the limit process X satisfies (2) with $\sigma(X_t) = \operatorname{diag}(\lambda_1(1 + c_{a,1}^2), \ldots, \lambda_m(1 + c_{a,m}^2))^{1/2}$. In the case of Poisson arrivals, we have $c_{a,i}^2 = 1$, and thus, $\sigma(X_t) = \operatorname{diag}(2\lambda_1, \ldots, 2\lambda_m)^{1/2}$.

Remark 7. This scaling is different from that used in Arapostathis et al. [6], Atar et al. [10], and Harrison and Zeevi [26], where the centering term uses $n\rho_i$ for the processes $X_i^n(t)$ and $Z_i^n(t)$. Here we use the prelimit parameters λ_i^n/μ_i^n together with the adjustment ϱ^n/m, which can be regarded as the reallocation of the safety staffing. Recall that when $\varrho^n > 0$ (and $\varrho > 0$), the condition in (58) is equivalent to the positive-square-root safety staffing rule (Whitt [36]). In addition, the diffusion-scaled process $\hat{X}^n$ converges to the limiting diffusion X with the drift given in (7). This follows from the standard martingale convergence technique in Pang et al. [31] using the representation of $\hat{X}^n$ in (60).

3.1.2. Scheduling Policies.

We define the space

$$\mathcal{Z}^n(x) := \big\{z \in \mathbb{Z}_+^m : z_i \leq x_i\,,\ \|z\|_1 = n \wedge \|x\|_1\big\}.$$

A scheduling policy is called (stationary) Markov if $Z^n(t) = z(X^n(t), S^n(t))$ for some function $z : \mathbb{Z}_+^m \times \mathbb{R}_+^m \to \mathcal{Z}^n(x)$, in which case we identify the policy with the function z. Let $S^n(t) = (S_1^n(t), \ldots, S_m^n(t))$, where $S_i^n(t)$ denotes the age process for class-i customers, defined by

$$S_i^n(t) := t - \frac{1}{\lambda_i^n} \sum_{j=1}^{A_i^n(t)} R_{ij}\,, \qquad t \geq 0.$$

Let

$$r_i^n(s_i) := \lambda_i^n \frac{f_i(\lambda_i^n s_i)}{1 - F_i(\lambda_i^n s_i)}, \quad s_i \ge 0, \tag{61}$$

denote the scaled hazard-rate function for the interarrival times of $A_i^n(t)$.

Under a Markov policy, the process (X^n, S^n) is Markov with extended generator

$$\begin{aligned}\mathcal{A}_z^n g(x,s) := \sum_{i\in\mathcal{I}} \frac{\partial g(x,s)}{\partial s_i} + \sum_{i\in\mathcal{I}} r_i^n(s_i)\big(g(x+e_i, s - s_i e_i) - g(x,s)\big) \\ + \sum_{i\in\mathcal{I}} \big(\mu_i^n z_i + \gamma_i^n q_i(x,z)\big)\big(g(x-e_i,s) - g(x,s)\big),\end{aligned} \tag{62}$$

for $g \in \mathcal{C}_b(\mathbb{R}^m \times \mathbb{R}^m)$ and $(x,s) \in \mathbb{Z}_+^m \times \mathbb{R}_+^m$. Here $q_i(x,z) = x_i - z_i$, and $e_i \in \mathbb{R}^m$ denotes the vector with the ith element equal to one and the rest of its elements equal to zero.

Let

$$\hat{x}_i^n(x) := \frac{1}{\sqrt{n}}\left(x_i - \frac{\lambda_i^n}{\mu_i^n}\right) - \frac{\varrho^n}{m}, \quad \hat{z}_i^n(x) := \frac{1}{\sqrt{n}}\left(z_i - \frac{\lambda_i^n}{\mu_i^n}\right) - \frac{\varrho^n}{m}, \text{ and } \hat{q}_i^n(x,z) := \frac{q_i(x,z)}{\sqrt{n}}. \tag{63}$$

We let $\mathfrak{X}^n$ denote the state space of the process $\hat{X}^n$. This is a countable subset of $\mathbb{R}^m$. Because $x \mapsto \hat{x}^n(x)$ is invertible, the set $\mathcal{Z}^n(x)$ can be equivalently written as a function of $\hat{x}^n$, and abusing the notation, we write this as $\mathcal{Z}^n(\hat{x}^n)$. In order to keep the notation simple, we often drop the superscript n from $\hat{x}^n$ when it is used to denote a generic element of $\mathfrak{X}^n$.

3.2. Results with Renewal Arrivals

The first main result is Theorem 5, which is the counterpart of Theorem 1 for the nth system. In order to state this theorem and demonstrate its proof, we need some additional notation, which we introduce next.

Let V be the function in (17) with μ^n replacing μ in its definition and parameters $\varepsilon > 0$ and $\theta \in (0,1)$. Let

$$\zeta_i^n(\tau) := \zeta_i(\lambda_i^n \tau), \quad \tau \ge 0, i \in \mathcal{I}.$$

In Theorem 4, we use the Lyapunov function $\mathcal{V}^n$ defined by

$$\mathcal{V}^n(\hat{x}, s) := G^n(\hat{x}, s) + V(\hat{x}), \quad (\hat{x}, s) \in \mathfrak{X}^n \times \mathbb{R}_+^m, \tag{64}$$

with

$$G^n(\hat{x}, s) := \sum_{i\in\mathcal{I}} \big(1 - \zeta_i^n(s_i)\big)\big(V(\hat{x} + n^{-1/2} e_i) - V(\hat{x})\big), \quad (\hat{x}, s) \in \mathfrak{X}^n \times \mathbb{R}_+^m.$$

By Assumption 1, for any fixed θ, we can choose $\tilde{\varepsilon}_0 = \tilde{\varepsilon}_0(\theta) > 0$ small enough that

$$\varepsilon \left| \sum_{i\in\mathcal{I}} \frac{1}{\mu_i^n} \big(1 - \zeta_i^n(s_i)\big)\big(\theta\psi'(-y_i) + \psi'(y_i)\big) \right| \le \frac{1}{2}, \quad \forall\, \varepsilon \le \tilde{\varepsilon}_0(\theta),\ \forall\, (y,s) \in \mathbb{R}^m \times \mathbb{R}_+^m,\ \forall n \in \mathbb{N}. \tag{65}$$

Then, provided that $\varepsilon \le \tilde{\varepsilon}_0(\theta)$, we have

$$\frac{1}{2} V(y) \le \mathcal{V}^n(y,s) \le \frac{3}{2} V(y). \tag{66}$$

We define

$$\widehat{V}^n(x) := V\big(\hat{x}^n(x)\big) \quad \text{and} \quad \widehat{G}^n(x,s) := \sum_{i\in\mathcal{I}} \big(1 - \zeta_i^n(s_i)\big)\big(V(\hat{x}^n(x+e_i)) - V(\hat{x}^n(x))\big) \tag{67}$$

for $x \in \mathbb{Z}_+^m$. Then the generator $\widehat{\mathcal{A}}_z^n$ of the diffusion-scaled state process $(\hat{X}^n, S^n)$ under a policy z takes the form

$$\widehat{\mathcal{A}}_z^n V(\hat{x}, s) = \mathcal{A}_z^n \widehat{V}^n(x,s) \quad \text{and} \quad \widehat{\mathcal{A}}_z^n G(\hat{x}, s) = \mathcal{A}_z^n \widehat{G}^n(x,s), \tag{68}$$

where $\mathcal{A}_z^n$ is as defined in (62).

We need to introduce some constants used in the results. First, for a function f on $\mathbb{R}^m$, if we define

$$\mathfrak{d}f(x;y) := f(x+y) - f(x),$$

it then follows by a repeated use of the mean value theorem that there exists a constant $\widehat{C}_1$ such that

$$\left|\mathfrak{d}\widehat{V}^n(x \pm e_j; \pm e_i) - \mathfrak{d}\widehat{V}^n(x; \pm e_i)\right| \le \frac{1}{n}\widehat{C}_1 \varepsilon(\varepsilon+\theta)\widehat{V}^n(x), \qquad \forall\, i,j \in \mathcal{I}, \tag{69}$$

and the same bound holds for $\left|\mathfrak{d}\widehat{V}^n(x;e_i) + \mathfrak{d}\widehat{V}^n(x;-e_i)\right|$. Also, by Assumption 1, (61), and the convergence of the parameters in (57), there exists a constant $\widehat{C}_0^n$ depending on n (implicitly through λ_i^n) such that

$$\sup_{n\in\mathbb{N}} \max_{i\in\mathcal{I}} \left(\frac{r_i^n(\tau)}{n} \vee \left(1 + \zeta_i^n(\tau)\right)\right) \le \widehat{C}_0^n, \qquad \forall\, \tau \ge 0. \tag{70}$$

We define

$$\widetilde{C}_0^n := m^2 \widehat{C}_0^n \widehat{C}_1\,, \qquad \widetilde{C}_1^n := \widehat{C}_1\left(m^2 \widehat{C}_0^n \mu_i^n + m(m-1)\left(\widehat{C}_0^n\right)^2 + \sum_{i\in\mathcal{I}} \frac{\lambda_i^n}{n}\right), \tag{71}$$

and

$$\theta_0(n) := \frac{1}{1 + (\beta^n{}_{\max} - 1)^+} \wedge \frac{1}{2\mu_{\max}^n\left(\widetilde{C}_0^n + \widehat{C}_1\right)} \wedge \frac{\varrho^n}{m}\left(m + 2\varrho^n + 4\left(\widetilde{C}_1^n + m\widehat{C}_1\,\widehat{C}_2^n + m\widehat{C}_3^n\right)\right)^{-1}. \tag{72}$$

Recall $\tilde{\varepsilon}_0(\theta)$ in (65). We are ready to state the first main result of this section.

Theorem 4. *We enforce Assumption 1, and in addition, we assume that the hazard-rate functions $\{h_i\}_{i\in\mathcal{I}}$ are bounded. Suppose that $\varrho^n > 0$. Then there exists a positive constant $C_0^n(\varepsilon)$ such that the function $\mathcal{V}^n$ in (64), with parameters $\theta = \theta_0(n)$ and any $\varepsilon < \theta_0(n) \wedge \tilde{\varepsilon}_0(\theta)$, satisfies*

$$\widehat{\mathcal{A}}_z^n \mathcal{V}^n(\hat{x}, s) \le C_0^n(\varepsilon) - \varepsilon\frac{\varrho^n}{3m}\mathcal{V}^n(\hat{x}, s), \qquad \forall\,(\hat{x}, s) \in \mathcal{X}^n \times \mathbb{R}_+^m\,, \forall\, z \in \mathcal{Z}^n(\hat{x}). \tag{73}$$

In particular, under any work-conserving stationary Markov policy, the process $(\hat{X}^n, S^n)$ is positive Harris recurrent, and $V(\hat{x})$ is integrable under its invariant probability distribution.

Remark 8. It is clear from the Foster–Lyapunov equation (73) that the stability result in Theorem 4 holds for all $n \in \mathbb{N}$ such that $\varrho^n > 0$, and the same applies for Theorem 6 and Corollary 2. We want to emphasize that this is an important by-product of the approach in this paper. One should compare it with Gamarnik and Stolyar [20, theorem 2], where stability is only stated as an asymptotic property, or, in other words, that it holds for all large enough n.

The convergence of the parameters in (57) implies that if the limiting value $\varrho = \lim_{n\to\infty} \varrho^n$ is positive, then $\theta_0(n)$ and $C_0(\varepsilon)$ can be selected independent of n in a manner that (73) holds for all sufficiently large n. Analogous conclusions can be drawn for Theorems 5 and 6 and Corollary 2, which appear later in this section.

The difference in the constant multiplying the Lyapunov function between (29) and (73) is only because of the bound in (66).

For the proof of the Theorem 4, we need the following result.

Lemma 3. *With $\widehat{\mathcal{V}}^n(x,s) := \widehat{G}^n(x,s) + \widehat{V}^n(x)$, we have the following inequality:*

$$\begin{aligned}\mathcal{A}_z^n\, \widehat{\mathcal{V}}^n(x,s) \le{} & \sum_{i\in\mathcal{I}}\left(\frac{\varrho^n\mu_i^n}{m} + \mu_i^n \hat{z}_i + \gamma_i^n \hat{q}_i(x,z)\right)\sqrt{n}\,\mathfrak{d}\widehat{V}^n(x;-e_i) \\ & + \varepsilon(\varepsilon+\theta)\frac{1}{\sqrt{n}}\widetilde{C}_0^n \sum_{i\in\mathcal{I}} \gamma_i^n \hat{q}_i(x,z)\widehat{V}^n(x) + \varepsilon(\varepsilon+\theta)\widetilde{C}_1^n \widehat{V}^n(x),\end{aligned} \tag{74}$$

with $\widetilde{C}_0^n$ and $\widetilde{C}_1^n$ as defined in (71).

Proof. Recall the definitions in (67), and note that

$$\widehat{G}_i^n(x,s) = \big(1-\zeta_i^n(s_i)\big)\eth\widehat{V}^n(x;e_i),$$

with $\widehat{V}^n$ as defined in (67). It follows by direct differentiation that

$$\frac{\mathrm{d}\zeta_i^n(\tau)}{\mathrm{d}\tau} - r_i^n(\tau)\zeta_i^n(\tau) = -\lambda_i^n, \quad \tau \ge 0. \tag{75}$$

Thus, using (69), (70), and (75) and noting that $\zeta_i^n(0)=1$, we obtain

$$\begin{aligned}
\mathcal{A}_z^n\widehat{G}_i^n(x,s) &= -\left(\frac{\mathrm{d}\zeta_i^n(s_i)}{\mathrm{d}s_i} + r_i^n(s_i)\big(1-\zeta_i^n(s_i)\big)\right)\eth\widehat{V}^n(x;e_i)\\
&\quad + r_i^n(s_i)\big(1-\zeta_i^n(s_i)\big)\sum_{j\neq i,\, i\in\mathcal{I}}\Big(\eth\widehat{V}^n(x+e_j;e_i) - \eth\widehat{V}^n(x;e_i)\Big)\\
&\quad - \big(\mu_i^n z_i + \gamma_i^n q_i(x,z)\big)\big(1-\zeta_i^n(s_i)\big)\sum_{j\in\mathcal{I}}\Big(\eth\widehat{V}^n(x-e_j;e_i) - \eth\widehat{V}^n(x;e_i)\Big)\\
&\le \big(\lambda_i^n - r_i^n(s_i)\big)\eth\widehat{V}^n(x;e_i) + (m-1)\Big(\widehat{C}_0^n\Big)^2\widehat{C}_1\varepsilon(\varepsilon+\theta)\widehat{V}^n(x)\\
&\qquad + \frac{m}{n}\widehat{C}_0^n\widehat{C}_1\varepsilon(\varepsilon+\theta)\big(\mu_i^n z_i + \gamma_i^n q_i(x,z)\big)\widehat{V}^n(x).
\end{aligned} \tag{76}$$

Also,

$$\mathcal{A}_z^n\widehat{V}^n(x) = \sum_{i\in\mathcal{I}} r_i^n(s_i)\eth\widehat{V}^n(x;e_i) + \sum_{i\in\mathcal{I}}\big(\mu_i^n z_i + \gamma_i^n q_i(x,z)\big)\eth\widehat{V}^n(x;-e_i). \tag{77}$$

Applying the identities

$$z_i = \sqrt{n}\hat{z}_i + \frac{\lambda_i^n}{\mu_i^n} + \sqrt{n}\frac{\varrho^n}{m} \quad \text{and} \quad q_i(x,z) = \sqrt{n}\hat{q}_i(x,z) \tag{78}$$

to (77), we obtain

$$\begin{aligned}
\mathcal{A}_z^n\,\widehat{V}^n(x) &= \sum_{i\in\mathcal{I}}\Big(r_i^n(s_i)\eth\widehat{V}^n(x;e_i) + \lambda_i^n\eth\widehat{V}^n(x;-e_i)\Big)\\
&\qquad + \sum_{i\in\mathcal{I}}\left(\frac{\varrho^n\mu_i^n}{m} + \mu_i^n\hat{z}_i + \gamma_i^n\hat{q}_i(x,z)\right)\sqrt{n}\,\eth\widehat{V}^n(x;-e_i).
\end{aligned} \tag{79}$$

Combining (76) and (79) and applying once more the estimate in (69) and the inequality $|z_i|\le n$, we deduce that

$$\begin{aligned}
\mathcal{A}_z^n\,\widehat{\mathcal{V}}^n(x,s) &\le \sum_{i\in\mathcal{I}}\left(\frac{\varrho^n\mu_i^n}{m} + \mu_i^n\hat{z}_i + \gamma_i^n\hat{q}_i(x,z)\right)\sqrt{n}\,\eth\widehat{V}^n(x;-e_i)\\
&\quad + \varepsilon(\varepsilon+\theta)\frac{m^2}{\sqrt{n}}\widehat{C}_0^n\widehat{C}_1\gamma_i^n\hat{q}_i(x,z)\widehat{V}^n(x)\\
&\quad + \varepsilon(\varepsilon+\theta)\widehat{C}_1\left(m^2\widehat{C}_0^n\mu_i^n + m(m-1)\Big(\widehat{C}_0^n\Big)^2 + \sum_{i\in\mathcal{I}}\frac{\lambda_i^n}{n}\right)\widehat{V}^n(x).
\end{aligned}$$

This completes the proof. □

Proof of Theorem 4. This proof relies on comparing the right-hand side of (74) with the drift inequalities in Lemma 1. First, we fix $n\in\mathbb{N}$, and as done earlier, we suppress the n-dependence of $\hat{x}_i^n$, $\hat{z}_i^n$, and $\hat{q}_i^n$ in the calculations, in the interest of simplifying the notation. It is clear from (57) and (63) that $\hat{q}_i \ge 0$ if $x_i \ge n$ or, equivalently, if

$$\hat{x}_i \ge \vartheta_n := \sqrt{n}\big(1-\rho_i^n\big) - \frac{\sqrt{n}}{m}\left(1 - \sum_{i=1}^m \frac{\lambda_i^n}{n\mu_i^n}\right) \ge 0.$$

If $\varepsilon \le 1$, then $\psi_\varepsilon(x-y) - \psi_\varepsilon(x) \le -\varepsilon\frac{y}{2}$ and $\psi(-x-y) - \psi(y) \le y$ for all $x \ge 0$ and $y \in [0,1]$ by Definition 1. Thus, if $\theta \in (0, 1/2]$ and $\varepsilon \in (0,1]$, then $V(x-y) \le V(x)$ for all $x \ge 0$ and $y \in [0,1]$. This, of course, implies, because $\theta_0(n) < 1/2$, that $\mathfrak{d}\widehat{V}^n(x;-e_i) \le 0$ if $\hat{x}_i \ge 0$. Thus, if we write

$$\begin{aligned}\left(\frac{\varrho^n\mu_i^n}{m} + \mu_i^n\hat{z}_i + \gamma_i^n\hat{q}_i(x,z)\right)&\sqrt{n}\,\mathfrak{d}\widehat{V}^n(x;-e_i)\\ &= \left(\frac{\varrho^n\mu_i^n}{m} + \mu_i^n\hat{z}_i + \gamma_i^n\hat{q}_i(x,z)\,1_{\{\hat{x}_i<\vartheta_n\}}\right)\sqrt{n}\,\mathfrak{d}\widehat{V}^n(x;-e_i)\\ &\quad + \gamma_i^n\hat{q}_i(x,z)\,1_{\{\hat{x}_i\ge\vartheta_n\}}\sqrt{n}\,\mathfrak{d}\widehat{V}^n(x;-e_i),\end{aligned} \tag{80}$$

then the second term on the right-hand side of (80) is negative. It is also clear that

$$\left|\frac{\varrho^n\mu_i^n}{m} + \mu_i^n\hat{z}_i + \gamma_i^n\hat{q}_i(x,z)\,1_{\{\hat{x}_i<\vartheta_n\}}\right| \le \widehat{C}_2^n\sqrt{n} \tag{81}$$

for some constant $\widehat{C}_2^n$ depending on the parameters.

Using the identity

$$\widehat{V}^n(x \pm e_i) - \widehat{V}^n(x) \mp \partial_{x_i}\widehat{V}^n(x) = \int_0^1 (1-t)\,\partial_{x_ix_i}\widehat{V}^n(x \pm te_i)\,\mathrm{d}t, \tag{82}$$

we deduce that

$$\left|\widehat{V}^n(x \pm e_i) - \widehat{V}^n(x) \mp \partial_{x_i}\widehat{V}^n(x)\right| \le \frac{1}{n}\varepsilon(\varepsilon+\theta)\,\widehat{C}_1\,\widehat{V}^n(x), \tag{83}$$

where we use a common constant to satisfy (69) and (83). Thus, by (80), (81), and (83) and using also the identity

$$\partial_{x_i}\widehat{V}^n(x) = \frac{1}{\sqrt{n}}\partial_{\hat{x}_i}V(\hat{x}),$$

we obtain

$$\begin{aligned}&\left(\frac{\varrho^n\mu_i^n}{m} + \mu_i^n\hat{z}_i + \gamma_i^n\hat{q}_i(x,z)\right)\sqrt{n}\,\mathfrak{d}\widehat{V}^n(x;-e_i)\\ &= -\left(\frac{\varrho^n\mu_i^n}{m} + \mu_i^n\hat{z}_i + \gamma_i^n\hat{q}_i(x,z)\,1_{\{\hat{x}_i<\vartheta_n\}}\right)\partial_{\hat{x}_i}V(\hat{x})\\ &+ \gamma_i^n\hat{q}_i(x,z)\,1_{\{\hat{x}_i\ge\vartheta_n\}}\sqrt{n}\,\mathfrak{d}\widehat{V}^n(x;-e_i) + \varepsilon(\varepsilon+\theta)\widehat{C}_1\,\widehat{C}_2^n\,\widehat{V}^n(x).\end{aligned} \tag{84}$$

Similarly, addressing the second term on the right-hand side of (74), we write

$$\frac{1}{\sqrt{n}}\widetilde{C}_0^n\gamma_i^n\hat{q}_i(x,z) \le \widehat{C}_3^n + \frac{1}{\sqrt{n}}\widetilde{C}_0^n\gamma_i^n\hat{q}_i(x,z)1_{\{\hat{x}_i\ge\vartheta_n\}} \tag{85}$$

for some constant $\widehat{C}_3^n$. Using (68), (84), and (85), we deduce from (74) that

$$\begin{aligned}\widehat{\mathcal{A}}_z^n\,\mathcal{V}^n(\hat{x},s) \le &-\sum_{i\in\mathcal{I}}\left(\frac{\varrho^n\mu_i^n}{m} + \mu_i^n\hat{z}_i + \gamma_i^n\hat{q}_i(\hat{x},\hat{z})\,1_{\{\hat{x}_i<\vartheta_n\}}\right)\partial_{\hat{x}_i}V(\hat{x}) + \varepsilon(\varepsilon+\theta)\left(\widetilde{C}_1^n + m\widehat{C}_1\,\widehat{C}_2^n + m\widehat{C}_3^n\right)V(\hat{x})\\ &+\sum_{i\in\mathcal{I}}\left(\sqrt{n}\,\mathfrak{d}\widehat{V}^n(x;-e_i) + \varepsilon(\varepsilon+\theta)\frac{1}{\sqrt{n}}\widetilde{C}_0^n\widehat{V}^n(x)\right)\gamma_i^n\hat{q}_i(\hat{x},\hat{z})1_{\{\hat{x}_i\ge\vartheta_n\}},\end{aligned} \tag{86}$$

where we express $\hat{q}$ as a function of $\hat{x}$ and $\hat{z}$, slightly abusing the notation.

We now turn to the drift inequalities in Lemma 1. It follows by (18) that there exist a constant and $C_0^n(\varepsilon)$ such that

$$\begin{aligned}&\sum_{i\in\mathcal{I}}\left(-\frac{\varrho^n\mu_i^n}{m} - \mu_i^n(\hat{x}_i - \langle e,\hat{x}\rangle^+u_i) - \gamma_i^n\langle e,\hat{x}\rangle^+u_i1_{\{\hat{x}_i<\vartheta_n\}}\right)\partial_{\hat{x}_i}V(\hat{x})\\ &+\varepsilon(\varepsilon+\theta)\left(\widetilde{C}_1^n + m\widehat{C}_1\,\widehat{C}_2^n + m\widehat{C}_3^n\right)V(\hat{x}) \le C_0^n(\varepsilon) - \varepsilon\frac{\varrho^n}{2m}V(\hat{x})\end{aligned} \tag{87}$$

for all $(\hat{x}, u) \in \mathbb{R}^m \times \Delta$ and for all $\varepsilon \in (0, \theta_0(n))$.

Consider the first sum in (87). If $\langle e, x\rangle \le n$, then $\hat{z} = \hat{x}$ by work conservation. By the scaling in (59) combined with (58), we have

$$\langle e, \hat{x}\rangle = \frac{1}{\sqrt{n}}(\langle e, x\rangle - n).$$

Thus, $\langle e, \hat{x}\rangle > 0$ if and only if $\langle e, x\rangle > n$. Similarly, $\langle e, z\rangle = n$ if and only if $\langle e, \hat{z}\rangle = 0$. By contrast, if $\langle e, x - z\rangle > 0$, then we can write $z = x - \langle e, x - z\rangle u$ for some $u \in \Delta$. Thus, $\hat{z} = \hat{x} - \langle e, \hat{x} - \hat{z}\rangle u = \hat{x} - \langle e, \hat{x}\rangle u$ because $\langle e, \hat{z}\rangle = 0$. We have thus established that for all $x \in \mathbb{R}^m_+$, we have

$$\hat{z} = \hat{x} - \langle e, \hat{x}\rangle^+ u \quad \text{and} \quad \hat{q}(\hat{x}, \hat{z}) = \langle e, \hat{x}\rangle^+ u \tag{88}$$

for some $u \in \Delta$. It then follows from (88) that the sum of the first two terms on the right-hand side of (86) has the bound on the right-hand side of (87).

Next, consider the last term in (86). By (82) and (83), we have

$$\mathfrak{d}\widehat{V}^n(x; -e_i) \le -\partial_{x_i}\widehat{V}^n(x) + \frac{1}{n}\varepsilon(\varepsilon + \theta)\widehat{C}_1\widehat{V}^n(x),$$

and $\partial_{x_i}\widehat{V}^n(x) = \frac{\varepsilon}{\sqrt{n}\mu_i^n}\widehat{V}^n(x)$ when $\hat{x}_i \ge \vartheta_n$. Thus,

$$\left(\sqrt{n}\,\mathfrak{d}\widehat{V}^n(x; -e_i) + \varepsilon(\varepsilon + \theta)\frac{1}{\sqrt{n}}\widetilde{C}_0^n\widehat{V}^n(x)\right)1_{\{\hat{x}_i \ge \vartheta_n\}} \le -\varepsilon\left(\frac{1}{\mu_i^n} - (\varepsilon + \theta)\frac{1}{\sqrt{n}}\left(\widetilde{C}_0^n + \widehat{C}_1\right)\right)\widehat{V}^n(x),$$

which is negative for all $\varepsilon < \theta = \theta_0(n)$ by the definition of θ_0 in (72). Thus, in view of (66), we have established the Foster–Lyapunov equation (73) as claimed.

The remaining conclusions of the theorem are straightforward in view of the fact that $\{S^n(t)\}_{t\ge 0}$ is positive Harris recurrent, as shown in Konstantopoulos and Last [28]. Because $\varepsilon < \tilde{\varepsilon}_0(\theta)$, (66) implies that $\mathcal{V}^n$ is bounded from below in $\mathbb{R}^m \times \mathbb{R}^m_+$. □

In the theorem that follows, we assume strictly positive abandonment rates for all classes, and we use the Lyapunov function

$$\mathcal{V}^n(\hat{x}^n, s) := \sum_{i\in\mathcal{I}}\left(1 - \zeta_i^n(s_i)\right)\left(\varphi^n\left(\hat{x}_i^n + n^{-1/2}\right) - \varphi^n\left(\hat{x}_i^n\right)\right) + \sum_{i\in\mathcal{I}}\frac{\varphi^n\left(\hat{x}_i^n\right)}{\mu_i^n}, \tag{89}$$

with

$$\varphi^n\left(y\right) := \tilde{\varepsilon}_0(\theta^n)\theta^n\psi\left(-y\right) + \tilde{\varepsilon}_0(\theta^n)\psi\left(y\right), \quad y \in \mathbb{R},$$

$\tilde{\varepsilon}_0$ as in (65), and

$$\theta^n = 1 \wedge \frac{\left(1 - \beta^n_{\mathsf{min}}\right) \vee \frac{1}{2}}{\beta^n_{\mathsf{max}}}, \quad \beta_i^n := \frac{\gamma_i^n}{\mu_i^n}, \quad i \in \mathcal{I}. \tag{90}$$

Theorem 5. *Grant Assumption 1, and in addition, assume that hazard-rate functions $\{h_i\}_{i\in\mathcal{I}}$ are locally bounded. Suppose that $\gamma_i^n > 0$ for all $i \in \mathcal{I}$. Then there exist positive constants $c_0(n)$ and $c_1(n)$ depending only on $n \in \mathbb{N}$ such that the function $\mathcal{V}^n$ in (89) satisfies*

$$\widehat{\mathcal{A}}_z^n\mathcal{V}^n(\hat{x}, s) \le c_0(n) - c_1(n)\mathcal{V}^n(\hat{x}, s), \qquad \forall\,(\hat{x}, s) \in \mathcal{X}^n \times \mathbb{R}^m_+,\ \forall\, z \in \mathcal{Z}^n(\hat{x}).$$

In particular, under any work-conserving stationary Markov policy, the process $(\hat{X}^n, S^n)$ is positive Harris recurrent, and $\|\hat{x}\|_1$ is integrable under its invariant probability distribution.

Proof. The proof mimics that of Theorem 4, also using Remark 6. The important difference here is that if we let $\hat{\varphi}^n(x_i) := \varphi^n(\hat{x}_i^n(x_i))$ and

$$\hat{\phi}_i^n(x, s) := \left(1 - \zeta_i^n(s_i)\right)\mathfrak{d}\hat{\varphi}^n(x_i; 1),$$

then, following the steps in (76), we obtain

$$\mathcal{A}_z^n\hat{\phi}_i^n(x,s) = \big(\lambda_i^n - r_i^n(s_i)\big)\eth\hat{\varphi}^n(x_i;1) - \big(\mu_i^n z_i + \gamma_i^n q_i(x,z)\big)\big(1-\zeta_i^n(s_i)\big)\big(\eth\phi^n(x_i-1;1) - \eth\phi^n(x_i;1)\big). \tag{91}$$

As a result, the terms corresponding to the second line of (76), for which the assumption that the hazard-rate functions are bounded was invoked, are not present in (91). The rest of the proof is the same. □

Without assuming that the abandonment rates are positive, but with $\varrho^n > 0$, we obtain uniform stability, that is, tightness of the invariant distributions. To establish this, we scale the Lyapunov function in (89) with a parameter $\varepsilon > 0$. More precisely, we define

$$\mathcal{V}_\varepsilon^n(\hat{x},s) := \sum_{i\in\mathcal{I}}\big(1-\zeta_i^n(s_i)\big)\eth\varphi_\varepsilon^n\big(\hat{x}_i;n^{-1/2}\big) + \sum_{i\in\mathcal{I}}\frac{\varphi_\varepsilon^n(\hat{x}_i)}{\mu_i^n}, \tag{92}$$

with

$$\varphi_\varepsilon^n(y) := \theta^n\varepsilon\psi(-y) + \psi_\varepsilon(y), \quad y\in\mathbb{R}.$$

The parameter θ^n depends on certain bounds that we review next. First, as we have seen in (69), there is a constant $\hat{C}_1$ such that

$$\Big|\eth\varphi_\varepsilon^n\big(x\pm n^{-1/2}e_j;\pm n^{-1/2}e_i\big) - \eth\varphi_\varepsilon^n\big(x;\pm n^{-1/2}e_i\big)\Big| \le \frac{1}{n}\hat{C}_1\varepsilon(\varepsilon+\theta), \qquad \forall\, i,j\in\mathcal{I}.$$

Let also $\hat{C}_0^n$ be a bound for $\|\max_i \zeta_i^n\|_\infty$. With $\widehat{C}_2^n$ the constant in (81), we define

$$\overline{C}_0^n := m^2\hat{C}_0^n\hat{C}_1 \quad \text{and} \quad \overline{C}_1^n := \hat{C}_1\left(m^2\hat{C}_0^n\mu_i^n + \sum_{i\in\mathcal{I}}\frac{\lambda_i^n}{n}\right).$$

Let θ^n be equal to the right-hand side of (72) after we replace $\widehat{C}_1$, $\widetilde{C}_1^n$, and $\widetilde{C}_2^n$ with $\hat{C}_1$, $\overline{C}_1^n$, and $\overline{C}_2^n$, respectively.

Theorem 6. *Grant* Assumption 1, *and in addition, assume that hazard-rate functions $\{h_i\}_{i\in\mathcal{I}}$ are locally bounded. Suppose that $\varrho^n > 0$. Then there exist a cube K and a constant C depending on ε, ϱ^n, and θ^n, defined previously, such that the function $\mathcal{V}_\varepsilon^n$ in* (92) *satisfies*

$$\widehat{\mathcal{A}}_z^n\mathcal{V}_\varepsilon^n(\hat{x},s) \le \varepsilon C 1_K(\hat{x}) - \varepsilon\frac{\varrho^n}{3m}, \qquad \forall\,(\hat{x},s)\in\mathfrak{X}^n\times\mathbb{R}_+^m,\ \forall\, z\in\mathcal{Z}^n(\hat{x}),$$

and for all $\varepsilon \le \theta^n$. In particular, under any work-conserving stationary Markov policy, the process $(\hat{X}^n, S^n)$ is positive Harris recurrent.

Proof. We follow the proofs of Lemma 3 and Theorem 4 to obtain the analogous inequality to (86). The result then follows by applying the drift inequality in (18) and using the definition of θ^n. □

3.3. Results with Poisson Arrivals

In this section, we specialize the results to a sequence of queuing models with Poisson arrivals with rates λ_i^n, $i\in\mathcal{I}$. Here, under a stationary Markov policy, the process $\{X^n(t)\}_{t\ge0}$ is Markov with extended generator

$$\mathcal{A}_z^n f(x) := \sum_{i\in\mathcal{I}}\lambda_i^n\big(f(x+e_i)-f(x)\big) + \sum_{i\in\mathcal{I}}\big(\mu_i^n z_i + \gamma_i^n q_i(x,z)\big)\big(f(x-e_i)-f(x)\big). \tag{93}$$

Define $\widehat{\mathcal{A}}_z^n$ analogously to (68). Mimicking the proof of Theorem 4, we deduce the following, which we state without proof.

Corollary 2. *Assume that the arrival processes are Poisson. Suppose that $\varrho^n > 0$. Then, for some $\theta = \theta(n) > 0$, there exist positive constants $\hat{\varepsilon}_0(n)$ and $C_0^n(\varepsilon)$ such that the function V in* (17) *satisfies*

$$\widehat{\mathcal{A}}_z^n V(\hat{x}) \le C_0^n(\varepsilon) - \varepsilon\frac{\varrho^n}{2m}V(\hat{x}), \qquad \forall\,\hat{x}\in\mathfrak{X}^n,\ \forall\, z\in\mathcal{Z}^n(\hat{x}),$$

and for all $\varepsilon \in (0, \hat{\varepsilon}_0(n))$. In particular, under any work-conserving stationary Markov policy, the process $\{\hat{X}^n(t)\}_{t\ge 0}$ is exponentially ergodic, and $V(\hat{x})$ is integrable under its invariant probability measure.

Remark 9. Let $\mathfrak{Z}^n_{\mathrm{sm}}$ denote the class of work-conserving stationary Markov policies for the process $\hat{X}^n(t)$. Suppose that $\varrho > 0$, and let $P^{n,z}_t$ and π^n_z denote the transition probability and the stationary distribution, respectively, of $\hat{X}^n(t)$ under a policy $z \in \mathfrak{Z}^n_{\mathrm{sm}}$. Then Corollary 2 implies that there exist positive constants γ and C_γ not depending on n or z such that

$$\left\| P^{n,z}_t(\hat{x}, \cdot\,) - \pi^n_z(\cdot) \right\|_V \le C_\gamma V(\hat{x})\, \mathrm{e}^{-\gamma t}, \qquad \forall\, \hat{x} \in \mathcal{X}^n,\ \forall\, t \ge 0. \tag{94}$$

Also,

$$\sup_{n\in\mathbb{N}}\ \sup_{z\in\mathfrak{Z}^n_{\mathrm{sm}}} \int_{\mathcal{X}^n} V(\hat{x})\, \pi^n_z(\mathrm{d}\hat{x}) < \infty.$$

If ν^n denotes the distribution of $\hat{X}^n(0)$, then (94) implies that

$$\left\| P^{n,z}_t(\nu^n, \cdot\,) - \pi^n_z(\cdot) \right\|_V \le C_\gamma \nu^n(V)\, \mathrm{e}^{-\gamma t}, \quad \forall\, t \ge 0, \tag{95}$$

where $P^{n,z}_t(\nu^n, \cdot\,) := \int_{\mathcal{X}^n} \nu^n(\mathrm{d}\hat{x}) P^{n,z}_t(\hat{x}, \cdot\,)$ and $\nu^n(V) := \int_{\mathcal{X}^n} V(\hat{x}) \nu^n(\mathrm{d}\hat{x})$. In particular, if $\hat{X}^n(0)$ is such that $\sup_{n\in\mathbb{N}} \nu^n(V) < \infty$, then the convergence in (95) is uniform over $z \in \mathfrak{Z}^n_{\mathrm{sm}}$ and $n \in \mathbb{N}$.

We also wish to remark that provided that the jobs do not abandon the queues, that is, $\Gamma = 0$, the hypothesis $\varrho^n > 0$ is sharp. In fact, there is a dichotomy. As shown in Corollary 2, if $\varrho^n > 0$, then $\{X^n(t)\}_{t\ge 0}$ is uniformly exponentially ergodic. Following, for example, the proof in Arapostathis et al. [9, theorem 3.3], one can show that if $\varrho^n < 0$ and jobs do not abandon the queues, then $\{X^n(t)\}_{t\ge 0}$ is transient under any Markov scheduling policy.

As explained in Gamarnik and Stolyar [20, p. 33], under positive abandonment in all classes, the invariant distribution of $\hat{X}^n$ cannot integrate a function of the form $\mathrm{e}^{\varepsilon|\hat{x}^2|}$ for $\varepsilon > 0$, although the invariant probability distribution of the limit diffusion has this property, as seen in Theorem 2. The technique in the proof of Theorem 4 stumbles in (83) because this bound is no longer valid for the function $\widetilde{V}$ of Theorem 2.

Nevertheless, we have the following improvement of Corollary 2 under positive abandonment in all classes.

Theorem 7. *Assume that the arrival processes are Poisson. Suppose that $\liminf_{n\to\infty} \gamma^n_i > 0$ for all $i \in \mathcal{I}$. Then there exist positive constants $\breve{\kappa}_0(\eta)$ and $\breve{\kappa}_1(\eta)$ such that the function*

$$\breve{V}^n(\hat{x}) := \exp\Bigl(\Phi^*_{\eta,\theta^n}(\hat{x})\Bigr) = \exp\bigl(\eta\theta^n\Psi(-\hat{x}) + \eta\Psi(\hat{x})\bigr),$$

with θ^n given by (90), satisfies

$$\widehat{\mathcal{A}}^n_z \breve{V}^n(\hat{x}) \le \breve{\kappa}_0(\eta) - \breve{\kappa}_1(\eta)\|\hat{x}\|_1 \breve{V}^n(\hat{x}), \qquad \forall\, (\hat{x}, z) \in \mathcal{X}^n \times \mathcal{Z}^n(\hat{x}),$$

and for all sufficiently large n. In particular, the function $\exp(\eta\|\hat{x}^n\|_1)$ is integrable under the stationary distribution of $\{\hat{X}^n(t)\}_{t\ge 0}$ for all $\eta > 0$ under any work-conserving stationary Markov scheduling policy.

Proof. Let $\widehat{\mathcal{V}}^n(x) := \breve{V}(\hat{x}^n(x))$. Applying the operator in (93) to $\widehat{\mathcal{V}}^n$ and using the analogous bound to (83),

$$\begin{aligned}
\mathcal{A}^n_z \widehat{\mathcal{V}}^n(x) \le \sum_{i\in\mathcal{I}} \biggl[&\lambda^n_i \left(\partial_{x_i} \widehat{\mathcal{V}}^n(x) + \frac{1}{n}\eta(1+\theta)\, \widehat{C}\, \widehat{\mathcal{V}}^n(x) \right) \\
&+ \bigl(\mu^n_i z_i + \gamma^n_i q_i(x,z)\bigr) \left(-\partial_{x_i} \widehat{\mathcal{V}}^n(x) + \frac{1}{n}\eta(1+\theta)\, \widehat{C}\, \widehat{\mathcal{V}}^n(x) \right) \biggr]
\end{aligned}$$

for some constant $\widehat{C}$. Using (78), we write this as

$$\begin{aligned}
\widehat{\mathcal{A}}^n_z \breve{V}(\hat{x}) \le\ &\sum_{i\in\mathcal{I}} \left(-\frac{\varrho^n \mu^n_i}{d} - \mu^n_i \hat{z}_i - \gamma^n_i \hat{q}_i(\hat{x}, \hat{z}) \right) \partial_{\hat{x}_i} \breve{V}(\hat{x}) \\
&+ \eta(1+\theta)\, \widehat{C} \sum_{i\in\mathcal{I}} \left(\frac{\lambda^n_i}{n} + \frac{\mu^n_i}{n} z_i + \frac{1}{\sqrt{n}} \gamma^n_i (\hat{x}_i - \hat{z}_i) \right) \breve{V}(\hat{x}).
\end{aligned} \tag{96}$$

Thus, using the drift inequality in Remark 6 to bound the first term on the right-hand side of (96) and noting that the coefficient of $\breve{V}$ on the second term on the right-hand side is of order $\frac{1}{\sqrt{n}}\|x\|_1$, we establish the result. □

We conclude with the analogous result to Corollary 1. We need the following notation:

$$\hat{\mathcal{I}}_1 := \left\{ i \in \mathcal{I} : \limsup_{n\to\infty} \frac{\gamma_i^n}{\mu_i^n} < 1 \right\}.$$

Theorem 8. *Assume that the arrival processes are Poisson. Suppose that* $\liminf_{n\to\infty} \varrho^n > 0$. *Then the function* $\exp(\eta \sum_{i\in\hat{\mathcal{I}}_1} \hat{x}_i^n)$ *is integrable under the invariant probability distribution of* $\{\hat{X}^n(t)\}_{t\ge0}$ *for all* $\eta > 0$ *and for all sufficiently large* n.

The proof closely mimics that of Theorem 4 and is therefore omitted.

References

[1] Aghajani R, Ramanan K (2016) The limit of stationary distributions of many-server queues in the Halfin–Whitt regime. *Math. Oper. Res.* 45(3):1016–1055.
[2] Aghajani R, Ramanan K (2019) Ergodicity of an SPDE associated with a many-server queue*Ann. Appl. Probabilities* 29(2):994–1045.
[3] Arapostathis A, Pang G (2016) Ergodic diffusion control of multiclass multi-pool networks in the Halfin-Whitt regime. *Ann. Appl. Probabilities* 26(5):3110–3153.
[4] Arapostathis A, Pang G (2018) Infinite-horizon average optimality of the N-network in the Halfin-Whitt regime. *Math. Oper. Res.* 43(3):838–866.
[5] Arapostathis A, Pang G (2019) Infinite horizon asymptotic average optimality for large-scale parallel server networks. *Stochastic Processing Appl.* 129(1):283–322.
[6] Arapostathis A, Biswas A, Pang G (2015) Ergodic control of multi-class $M/M/N+M$ queues in the Halfin-Whitt regime. *Ann. Appl. Probabilities* 25(6):3511–3570.
[7] Arapostathis A, Borkar VS, Ghosh MK (2011) Ergodic control of diffusion processes. *Encyclopedia of Mathematics and its Applications*, vol. 143 (Cambridge University Press, Cambridge, UK).
[8] Arapostathis A, Borkar VS, Kumar KS (2014) Convergence of the relative value iteration for the ergodic control problem of nondegenerate diffusions under near-monotone costs. *SIAM J. Control Optim.* 52(1):1–31.
[9] Arapostathis A, Pang G, Sandrić N (2019) Ergodicity of a Lévy-driven SDE arising from multiclass many-server queues. *Ann. Appl. Probabilities* 29(2):1070–1126.
[10] Atar R, Mandelbaum A, Reiman MI (2004) Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probabilities* 14(3):1084–1134.
[11] Braverman A, Dai JG (2017) Stein's method for steady-state diffusion approximations of $M/\mathrm{Ph}/n+M$ systems. *Ann. Appl. Probabilities* 27(1):550–581.
[12] Braverman A, Dai JG, Feng J (2016) Stein's method for steady-state diffusion approximations: An introduction through the Erlang-A and Erlang-C models. *Stochastic Systems* 6(2):301–366.
[13] Braverman A, Dai JG, Miyazawa M (2017) Heavy traffic approximation for the stationary distribution of a generalized Jackson network: The BAR approach. *Stochastic Systems* 7(1):143–196.
[14] Budhiraja A, Lee C (2009) Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Math. Oper. Res.* 34(1):45–56.
[15] Dai JG, He S (2013) Many-server queues with customer abandonment: Numerical analysis of their diffusion model. *Stochastic Systems* 3(1):96–146.
[16] Dai JG, Dieker AB, Gao X (2014) Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. *Queueing Systems* 78(1):1–29.
[17] Dieker AB, Gao X (2013) Positive recurrence of piecewise Ornstein-Uhlenbeck processes and common quadratic Lyapunov functions. *Ann. Appl. Probabilities* 23(4):1291–1317.
[18] Gamarnik D, Goldberg DA (2013) Steady-state GI/GI/n queue in the Halfin-Whitt regime. *Ann. Appl. Probabilities* 23(6):2382–2419.
[19] Gamarnik D, Momčilović P (2008) Steady-state analysis of a multiserver queue in the Halfin-Whitt regime. *Adv. Appl. Probabilities* 40(2):548–577.
[20] Gamarnik D, Stolyar AL (2012) Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime: asymptotics of the stationary distribution. *Queueing Systems* 71(1–2):25–51.
[21] Gamarnik D, Zeevi A (2006) Validity of heavy traffic steady-state approximation in generalized Jackson networks. *Ann. Appl. Probabilities* 16(1):56–90.
[22] Gruber M (1984) Harnack inequalities for solutions of general second order parabolic equations and estimates of their Hölder constants. *Math. Zeitschrift* 185(1):23–43.
[23] Gurvich I (2014) Validity of heavy-traffic steady-state approximations in multiclass queueing networks: The case of queue-ratio disciplines. *Math. Oper. Res.* 39(1):121–162.
[24] Gyöngy I, Krylov N (1996) Existence of strong solutions for Itô's stochastic equations via approximations. *Probabilities Theory Related Fields* 105(2):143–158.
[25] Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.
[26] Harrison JM, Zeevi A (2004) Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Oper. Res.* 52(2):243–257.
[27] Katsuda T (2010) State-space collapse in stationarity and its application to a multiclass single-server queue in heavy traffic. *Queueing Systems* 65(3):237–273.

[28] Konstantopoulos T, Last G (1999) On the use of Lyapunov function methods in renewal theory. *Stochastic Processing Appl.* 79(1):165–178.
[29] Meyn SP, Tweedie RL (1993) Stability of Markovian processes. III. Foster-Lyapunov criteria for continuous-time processes. *Adv. Appl. Probabilities* 25(3):518–548.
[30] Meyn SP, Tweedie RL (1994) Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probabilities* 4(4):981–1011.
[31] Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probabilities Survey* 4:193–267.
[32] Stolyar AL (2015) Diffusion-scale tightness of invariant distributions of a large-scale flexible service system. *Adv. Appl. Probabilities* 47(1):251–269.
[33] Stolyar AL (2015) Tightness of stationary distributions of a flexible-server system in the Halfin-Whitt asymptotic regime. *Stochastic Systems* 5(2):239–267.
[34] Stolyar AL, Yudovina E (2012) Tightness of invariant distributions of a large-scale flexible service system under a priority discipline. *Stochastic Systems* 2(2):381–408.
[35] Stolyar AL, Yudovina E (2013) Systems with large flexible server pools: Instability of "natural" load balancing. *Ann. Appl. Probabilities* 23(5):2099–2138.
[36] Whitt W (1992) Understanding the efficiency of multi-server service systems. *Management Sci.* 38(5):708–723.
[37] Ye HQ, Yao DD (2016) Diffusion limit of fair resource control: Stationarity and interchange of limits. *Math. Oper. Res.* 41(4):1161–1207.
[38] Ye HQ, Yao DD (2018) Justifying diffusion approximations for multiclass queueing networks under a moment condition. *Ann. Appl. Probabilities* 28(6):3652–3697.