Uniform stability of some large-scale parallel server networks

HASSAN HMEDI*, ARI ARAPOSTATHIS*, AND GUODONG PANG†

Abstract. In this paper we study the uniform stability properties of two classes of parallel server networks with multiple classes of jobs and multiple server pools of a tree topology. These include a class of networks with a single non-leaf server pool, such as the 'N' and 'M' models, and networks of any tree topology with class-dependent service rates. We show that with $\sqrt{n}$ safety staffing, and no abandonment, in the Halfin–Whitt regime, the diffusion-scaled controlled queueing processes are exponentially ergodic and their invariant probability distributions are tight, *uniformly* over all stationary Markov controls. We use a unified approach in which the same Lyapunov function is used in the study of the prelimit and diffusion limit.

A parameter called *the spare capacity (safety staffing) of the network* plays a central role in characterizing the stability results: the parameter being positive is necessary and sufficient that the limiting diffusion is uniformly exponentially ergodic over all stationary Markov controls. We introduce the concept of "*system-wide work conserving policies*", which are defined as policies that minimize the number of idle servers at all times. This is stronger than the so-called joint work conservation. We show that, provided the spare capacity parameter is positive, the diffusion-scaled processes are geometrically ergodic and the invariant distributions are tight, uniformly over all "system-wide work conserving policies". In addition, when the spare capacity is negative we show that the diffusion-scaled processes are transient under any stationary Markov control, and when it is zero, they cannot be positive recurrent.

1. Introduction

Large-scale parallel server networks have been the subject of intense study, due to their use in modeling a variety of systems including telecommunications, data centers, customer services and manufacturing systems; see, e.g., [1, 9, 14, 19, 20, 25, 30, 31]. In such networks, there are multiple classes of jobs and multiple server pools where each job class can be served by a subset of server pools while each server pool can serve a subset of job classes, thus requiring optimal routing and scheduling decisions. Many of these systems operate in the so-called the Halfin–Whitt regime (or Quality-and-Efficiency-Driven (QED) regime [12, 23, 32]), where the arrival rates and the numbers of servers grow large as the scale of the system grows, while the service rates remain fixed in such a way that the system becomes critically loaded.

Ensuring stability of these systems through allocating available resources by means of adjusting controller parameters is of great importance. Existing work in the literature has addressed the following important questions:

(i) Uniform stability of the multiclass single-pool "V" network. The study in [17] focused on the prelimit diffusion-scaled process and showed that, with square-root safety staffing in the single-pool of servers, the invariant probability distributions under all work-conserving scheduling policies are tight, and have a uniform exponential tail when the model has no

* Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712

† Department of Computational Applied Mathematics and Operations Research, George R. Brown College of Engineering, Rice University, Houston, TX 77005

E-mail addresses: {hmedi,ari}@utexas.edu, gdpang@rice.edu.

2000 *Mathematics Subject Classification.* Primary: 90B22. Secondary: 60K25; 49L20; 90B36.

Key words and phrases. uniform exponential ergodicity, parallel server (multiclass multi-pool) networks, Halfin–Whitt regime, spare capacity, system-wide work conservation.

abandonment (or a sub-Gaussian tail with abandonment). In [4], a unified approach with a common Lyapunov function is developed to establish a Foster-Lyapunov equation for both the diffusion limit and the diffusion-scaled processes, which shows that the associated invariant probability measures have exponential tails, uniformly over the scale of the network, and over all stationary (work-conserving) Markov controls.

(ii) Stability of the 'N' network under a static priority scheduling policy. With safety staffing in one server pool and no abandonment, Stolyar [27] employed a integral type of Lyapunov function and established the tightness of stationary distributions of the diffusion-scaled process (there is no analysis of the rate of convergence though).

(iii) Counterexamples for stability of multi-class multi-pool networks. Stolyar and Yudovina [29] showed that the stationary distributions of the diffusion-scaled processes may not be tight in these regimes under a natural load balancing scheduling policy, "Longest-queue freest-server" (LQFS-LB) (also true in the underloaded regime).

(iv) Stability of multi-class multi-pool networks with pool-dependent service rates under the LQFS-LB policy [29]. We also refer the reader to [26, 28], even though these concern the underloaded case.

(v) Stability of multiclass multi-pool networks under a family of Markov policies. In [6, 7], it is shown that a class of state-dependent policies, referred to as *balanced saturation policies* (BSP) are stabilizing for the prelimit diffusion-scaled queueing process, when at least one abandonment rate is strictly positive.

(vi) Stability of the limiting controlled diffusions for multiclass multi-pool networks under a constant control. Arapostathis and Pang [5] developed a leaf elimination algorithm to derive an explicit expression of the drift, and, consequently, by using the structural properties of the drift, a static priority scheduling and routing control is identified which stabilizes the limiting diffusion, when at least one of the classes has a positive abandonment rate.

(vii) Stabilizability of multiclass multi-pool networks of any tree topology without abandonment in the Halfin-Whitt regime. Hmedi, Arapostahis and Pang [24] identified a system-wide safety staffing parameter and showed that that parameter being positive is a necessary and sufficient condition for the network to be stabilizable, that is, there exists a scheduling policy under which the stationary distributions of the controlled diffusion-scaled queueing processes are tight over the size of the network.

The stability results in (v) and (vi) are used in the aforementioned papers for the study of ergodic control problems for multiclass multi-pool networks. In [2, 5–7], due to the lack of the "uniform stability" (also called "blanket stability") property, ergodic control problems were studied using a rather elaborate methodology. The uniform stability properties established in this paper render the ergodic control problem much simpler, and it can be studied by applying the methodology in [3, Chapter 3.7].

Despite all the important results in (i)–(vi), the ergodic properties of multiclass multi-pool networks in the Halfin–Whitt regime are far from being well understood. The stability analysis of multiclass multi-pool networks in the Halfin–Whitt regime is considerably more challenging than the corresponding one for the 'V' network. The problem is particularly difficult when the system does not have abandonment.

Given the counterexamples in [29], uniform stability, that is, tightness of the invariant probability distributions, does not hold for multiclass multi-pool networks of any tree topology. In this paper we identify a large class of such networks that are indeed uniformly stable: (a) networks with one dominant server pool, that is, a single non-leaf server pool, which include the 'N', 'M' and generalized 'N', 'M' networks with diameters equal to three or four, and (b) networks with class-dependent service rates. It might appear to the reader that the topology in (a) is restrictive. One should note though that even for simple networks with two non-leaf server pools [29, Figure 2, p. 21] the parameters can be chosen so that uniform stability fails.

The classes of networks in (a)–(b) share an important structural property in their drift, that is, the matrix B_1 in (2.30) is diagonal. We establish a necessary and sufficient condition for uniform stability, via the so-called spare capacity (safety staffing) parameter defined in (2.35). For the networks under consideration, we show that if the spare capacity is negative, then the limiting diffusion is transient under any stationary Markov control, if it is zero, the diffusion cannot be positive recurrent, and if it is positive, the diffusion limit is uniformly stable over all stationary Markov controls (see Proposition 3.1 and Theorem 2.1). The analogous results for the diffusion-scaled processes are also established. Lastly, we provide a characterization of the spare capacity parameter for the limiting diffusion when the latter is positive recurrent. We show in Theorem 3.1 that the spare capacity is equal to an average 'idleness' weighted by the critical quantity in (3.1).

To prove the uniform exponential ergodicity for the limiting controlled diffusion, we use a common Lyapunov function given in (4.7). This Lyapunov function consists of two components that treat the positive and negative half spaces of the state space in a delicate manner. An important 'tilting' parameter must be carefully chosen to account for not only the different effects of queueing and idleness (positive and negative half state space), but also the second order derivatives of the extended generator of the diffusion. Note that these Lyapunov functions differ from the quadratic Lyapunov functions used in [5–8, 16] for the study of stability under either constant controls or assuming abandonment, and also differ from that used in [4] for the uniform stability of the 'V' network. In [16] for example, the stability analysis requires the existence of a common quadratic Lyapunov function which cannot be shown for the models under consideration. As it will be clear to the reader later in the paper, the uniform stability analysis without abandonment requires the use of the sum of two functions where each of them 'dominates' the other over a part of the state space. See for example the proofs of Lemmas 4.1 and 4.2.

The same Lyapunov function is used to prove the uniform exponential ergodicity for the prelimit diffusion-scaled processes. However, unlike the 'V' network studied in [4], the Foster-Lyapunov equations for the limiting diffusion do not carry over to the analogous equations for the diffusion-scaled queueing processes over the entire state space. The reason lies in the jointly work conserving (JWC) condition (that is, all the queues have to be empty when there are idle servers) which is essential in establishing the weak convergence to the controlled limiting diffusion (see [10, 11]). To tackle this difficulty, we first provide an explicit 'drift' representation of the diffusion-scaled processes which differs from the drift of the diffusion by an extra term that accounts for the deviation from the JWC condition in the n^{th} system, and which vanishes in the limit. A natural extension of the concept of work conservation for multiclass multi-pool networks is minimization of the idle servers at all times. This defines an action space which we call *system-wide work conserving* (SWC). Establishing the "uniform" geometric ergodicity over all SWC Markov policies when the spare capacity is positive, is accomplished by first proving a useful upper bound for the minimum of idle servers and cumulative queue size for the n^{th} system, and then using this to derive the Foster–Lyapunov drift inequalities in the region of the state space where the drifts of the diffusion limit and the n^{th} system do not match. This facilitates establishing the drift inequalities for the diffusion-scaled processes. As a consequence of the Foster-Lyapunov equations, the invariant probability measures of the diffusion-scaled queueing processes have uniform exponential tails.

The property of interchange of limits attests to the validity of the diffusion approximation for the queueing network. For stochastic networks in the conventional heavy traffic regime, we refer the readers to the papers [13, 15, 18, 21, 34, 35] and references therein. For the 'V' network in the Halfin–Whitt regime, interchange of limits is established in [4, 17]. For the 'N' network, Stolyar [27] has shown the interchange of limits under a specific static priority policy. This property also holds for networks with pool-dependent service rates under the LQFS-LB scheduling policy, as shown in [29, Section 7.2]. Stolyar and Yudovina [28] and Stolyar [27] then proved tightness of the stationary distributions and interchange of limits of a leaf-activity priority policy in the sub-diffusion and diffusion scales, respectively, in the underloaded regime. This paper contributes to this literature by establishing that the limit of the diffusion-scaled invariant distributions is equal to the invariant distribution of the limiting diffusion process for the large classes of networks considered under any stationary Markov policy (see Remark 5.3).

1.1. **Organization of the paper.** In the next subsection, we summarize the notation used in the paper. In Subsection 2.1, we describe the model and state informally the assumptions used. We define the diffusion scaled processes, and characterize the corresponding controlled generator in Subsection 2.2. In Subsection 2.3, the notion of *system-wide work conserving policies* is introduced, and this is used in Subsection 2.4 to take limits and establish the diffusion approximation. In Section 3, we define the parameter of *spare capacity* (ϱ) for multiclass multi-pool networks and show that whenever $\varrho < 0$, the process is transient under any stationary Markov control both for the diffusion limit and the n^{th} system for the models under consideration. In the same subsection, we establish the relation between the spare capacity and average idleness. In Section 4 we first provide equivalent characterizations of uniform exponential ergodicity of controlled diffusions, and then proceed to establish that the diffusion limits of the aforementioned classes of networks are uniformly exponentially ergodic and their invariant probability measures have uniform exponential tails. Finally, Section 5 is devoted to the study of uniform exponential ergodicity of the n^{th} system of networks under consideration.

1.2. **Notation.** 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 the real line. We use z^{T} to denote the transpose of a vector $z \in \mathbb{R}^m$. Throughout the paper $e \in \mathbb{R}^m$ stands for the vector whose elements are equal to 1, that is, $e = (1, \dots, 1)^{\mathsf{T}}$, and $e_i \in \mathbb{R}^m$ denotes the vector whose elements are all 0 except for the i^{th} element which is equal to 1. For $x, y \in \mathbb{R}$, $x \vee y = \max\{x, y\}$, $x \wedge y = \min\{x, y\}$, $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$.

For a set $A \subseteq \mathbb{R}^m$, we use A^{c}, ∂A, and $\mathbb{1}_A$ 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. For $x \in \mathbb{R}^m$, we let $\|x\|_1 := \sum_i |x_i|$, and by K_r, or $K(r)$, for $r > 0$, we denote the closed cube

$$K_r := \{x \in \mathbb{R}^m \colon \|x\|_1 \leq r\}\,. \tag{1.1}$$

Also, we define $x_{\mathsf{max}} := \max_i\, x_i$, and $x_{\mathsf{min}} := \min_i\, x_i$, and $x^{\pm} := \bigl(x_1^{\pm}, \dots, x_m^{\pm}\bigr)$.

For a finite signed measure ν on $\mathbb{R}^m$, and a Borel measurable $f \colon \mathbb{R}^m \to [1, \infty)$, the f-norm of ν is defined by

$$\|\nu\|_f := \sup_{g \in \mathcal{B}(\mathbb{R}^m),\ |g| \leq f} \left| \int_{\mathbb{R}^m} g(x)\, \nu(\mathrm{d}x) \right|, \tag{1.2}$$

where $\mathcal{B}(\mathbb{R}^m)$ denotes the class of Borel measurable functions on $\mathbb{R}^m$.

2. The queueing network model and the diffusion limit

In this section, we consider a sequence of parallel server networks whose processes, parameters, and variables are indexed by n. We recall some of the definitions and notations used in [5,7].

2.1. **Model and assumptions.** Consider a general Markovian parallel server (multiclass multi-pool) network with m classes of customers and J server pools. Customer classes take values in $\mathcal{I} = \{1, \dots, m\}$ and server pools in $\mathcal{J} = \{1, \dots, J\}$. Forming their own queue, customers of each class are served according to a First-Come-First-Served (FCFS) service discipline. We assume throughout the paper that customers do not abandon. For all $i \in \mathcal{I}$, let $\mathcal{J}(i)$ denote the subset of server pools that can serve customer class i. On the other hand, for all $j \in \mathcal{J}$, let $\mathcal{I}(j)$ be the subset of customer classes that can be served by server pool j.

We form a bipartite undirected graph $\mathcal{G} = (\mathcal{I} \cup \mathcal{J}, \mathcal{E})$ with a set of edges defined by $\mathcal{E} = \{(i,j) \in \mathcal{I} \times \mathcal{J} \colon j \in \mathcal{J}(i)\}$, and use the notation $i \sim j$, if $(i,j) \in \mathcal{E}$, and $i \nsim j$, otherwise. We assume that the graph $\mathcal{G}$ is a tree. We define

$$\mathbb{R}^{\mathcal{G}}_{+} \coloneqq \left\{\xi = [\xi_{ij}] \in \mathbb{R}^{m\times J}_{+} : \xi_{ij} = 0 \text{ for } i \nsim j\right\}, \tag{2.1}$$

and analogously define $\mathbb{R}^{\mathcal{G}}$, $\mathbb{Z}^{\mathcal{G}}_{+}$, and $\mathbb{Z}^{\mathcal{G}}$.

In each server pool j, we let N^n_j be the number of servers, and assume that the servers are statistically identical. For each $i \in \mathcal{I}$, class i customer arrives according to a Poisson process with arrival rate $\lambda^n_i > 0$. These customers are served at an exponential rate $\mu^n_{ij} > 0$ at server pool j if $j \in \mathcal{J}(i)$, and $\mu^n_{ij} = 0$ otherwise. Finally, we assume that the arrival and service processes of all classes are mutually independent. We study these networks in the Halfin–Whitt regime, which involves the following assumption on the parameters. There exist positive constants λ_i and ν_j, nonnegative constants μ_{ij}, with $\mu_{ij} > 0$ for $i \sim j$ and $\mu_{ij} = 0$ for $i \nsim j$, and constants $\hat{\lambda}_i$, $\hat{\mu}_{ij}$ and $\hat{\nu}_j$, such that the following limits exist as $n \to \infty$:

$$\frac{\lambda^n_i - n\lambda_i}{\sqrt{n}} \;\to\; \hat{\lambda}_i\,, \qquad \sqrt{n}\,(\mu^n_{ij} - \mu_{ij}) \;\to\; \hat{\mu}_{ij}\,, \quad \text{and} \quad \frac{N^n_j - n\nu_j}{\sqrt{n}} \;\to\; \hat{\nu}_j\,. \tag{2.2}$$

The parameters λ_i and μ_{ij} are the limiting arrival and service rates, ν_j is the limiting service capacity in pool j in the fluid scale, while the parameters $\hat{\lambda}_i$, $\hat{\mu}_{ij}$ and $\hat{\nu}_j$ are the associated limits in the diffusion scale.

An additional standard assumption referred to as the *complete resource pooling* condition [11,33] concerns the fluid scale equilibrium, and is stated as follows. The linear program (LP) given by

$$\text{Minimize} \quad \max_{j\in\mathcal{J}} \sum_{i\in\mathcal{I}(j)} \xi_{ij}\,, \quad \text{subject to} \quad \sum_{j\in\mathcal{J}(i)} \mu_{ij}\nu_j\xi_{ij} \;=\; \lambda_i \;\;\forall\, i\in\mathcal{I}\,, \tag{2.3}$$

has a unique solution $\xi^* = [\xi^*_{ij}] \in \mathbb{R}^{\mathcal{G}}_{+}$ satisfying

$$\sum_{i\in\mathcal{I}} \xi^*_{ij} \;=\; 1, \quad \forall j\in\mathcal{J}\,, \quad \text{and} \quad \xi^*_{ij} > 0 \quad \text{for all } i\sim j\,. \tag{2.4}$$

We define $x^* \in \mathbb{R}^m$, and $z^* \in \mathbb{R}^{\mathcal{G}}_{+}$ by

$$x^*_i = \sum_{j\in\mathcal{J}} \xi^*_{ij}\nu_j\,, \quad \text{and} \quad z^*_{ij} = \xi^*_{ij}\nu_j\,. \tag{2.5}$$

The quantity ξ^*_{ij} represents the fraction of servers in pool j allocated to class i in the fluid equilibrium, x^*_i represents the total number of class i customers in the system, and z^*_{ij} is the number of class i customers in pool j. Note that the constraint in (2.3) is the rate balance equation for each class i with allocations in each service pool j. Also, $\rho_j \coloneqq \sum_i \xi_{ij}$ can be interpreted as the traffic intensity in pool j, hence the condition in (2.4) implies that each pool is critically loaded.

For each $i \in \mathcal{I}$ and $j \in \mathcal{J}$, we let $X^n_i = \{X^n_i(t) \colon t \geq 0\}$ denote the total number of class i customers in the system (both in service and in queue), $Z^n_{ij} = \{Z^n_{ij}(t),\, t \geq 0\}$ the number of class i customers currently being served in pool j, $Q^n_i = \{Q^n_i(t),\, t\geq 0\}$ the number of class i customers in the queue, and $Y^n_j = \{Y^n_j(t),\, t \geq 0\}$ the number of idle servers in server pool j. Let $X^n = (X^n_i)_{i\in\mathcal{I}}$, $Y^n = (Y^n_j)_{j\in\mathcal{J}}$, $Q^n = (Q^n_i)_{i\in\mathcal{I}}$, and $Z^n = (Z^n_{ij})_{i\in\mathcal{I},\,j\in\mathcal{J}}$. The process Z^n is the scheduling control. We have clearly the following *balance equations*

$$\begin{aligned} Q^n_i(t) \;&\coloneqq\; X^n_i(t) - \sum_{j\in\mathcal{J}} Z^n_{ij}(t)\,, \quad i\in\mathcal{I}\,, \\ Y^n_j(t) \;&\coloneqq\; N^n_j - \sum_{i\in\mathcal{J}} Z^n_{ij}(t)\,, \quad j\in\mathcal{J}\,, \end{aligned} \tag{2.6}$$

Dropping the explicit dependence on n for simplicity, let $(x, z) \in \mathbb{Z}_+^m \times \mathbb{Z}_+^{\mathcal{G}}$ denote a state-action pair. We rewrite (2.6) as

$$\begin{aligned} q_i(x,z) &:= x_i - \sum_{j\in\mathcal{J}} z_{ij}\,, \quad i \in \mathcal{I}\,, \\ y_j(z) &:= N_j^n - \sum_{i\in\mathcal{J}} z_{ij}\,, \quad j \in \mathcal{J}\,. \end{aligned} \tag{2.7}$$

and define the *action space* $\mathcal{Z}^n(x)$ by

$$\mathcal{Z}^n(x) := \bigl\{z \in \mathbb{Z}_+^{\mathcal{G}} : q_i(x,z) \wedge y_j^n(z) = 0\,,\ q_i(x,z) \ge 0\,,\ y_j^n(z) \ge 0 \quad \forall\,(i,j) \in \mathcal{E}\bigr\}\,.$$

Note that this space consists of work-conserving actions only. It should be noted here that there is an abuse of notation in (2.7). The quantities $q_i(x, z)$ and $y_j(z)$ still represent the number of class i customers in the queue and the number of idle servers in pool j respectively. Equation (2.7) is used to show the dependence on x and z through the balance equations.

2.2. **Diffusion scaling.** With $\xi^* \in \mathbb{R}_+^{\mathcal{G}}$ the solution of the (LP), we define the centering quantities of the diffusion-scaled processes $\bar{z}^n \in \mathbb{R}_+^{\mathcal{G}}$ and $\bar{x}^n \in \mathbb{R}^m$ by

$$\bar{z}_{ij}^n := \frac{1}{n}\xi_{ij}^* N_j^n\,, \qquad \bar{x}_i^n := \sum_{j\in\mathcal{J}} \bar{z}_{ij}^n\,, \tag{2.8}$$

and

$$\begin{aligned} \hat{X}_i^n(t) &:= \frac{1}{\sqrt{n}}\bigl(X_i^n(t) - n\bar{x}_i^n\bigr)\,, \qquad & \hat{Z}_{ij}^n(t) &:= \frac{1}{\sqrt{n}}\bigl(Z_{ij}^n(t) - n\bar{z}_{ij}^n\bigr)\,, \\ \hat{Q}_i^n(t) &:= \frac{1}{\sqrt{n}} Q_i^n(t)\,, & \hat{Y}_j^n(t) &:= \frac{1}{\sqrt{n}} Y_j^n(t)\,. \end{aligned} \tag{2.9}$$

Using (2.6) and (2.9), these obey the *centered balance equations*

$$\begin{aligned} \hat{X}_i^n(t) &= \hat{Q}_i^n(t) + \sum_{j\in\mathcal{J}(i)} \hat{Z}_{ij}^n(t) \qquad \forall\, i \in \mathcal{I}\,, \\ \hat{Y}_j^n(t) + \sum_{i\in\mathcal{I}(j)} \hat{Z}_{ij}^n(t) &= 0 \qquad \forall\, j \in \mathcal{J}\,. \end{aligned} \tag{2.10}$$

We introduce suitable notation in the diffusion scale as follows (see [7, Definition 2.3]).

For $x \in \mathbb{Z}_+^m$ and $z \in \mathcal{Z}^n(x)$, we define

$$\hat{x}^n := \frac{x - n\bar{x}^n}{\sqrt{n}}\,, \qquad \hat{z}^n := \frac{z - n\bar{z}^n}{\sqrt{n}}\,, \tag{2.11}$$

and let $\mathcal{S}^n$ denote the state space in the diffusion scale, that is,

$$\mathcal{S}^n := \bigl\{\hat{x} \in \mathbb{R}^m : \sqrt{n}\hat{x} + n\bar{x}^n \in \mathbb{Z}_+^m\bigr\}\,. \tag{2.12}$$

It is clear that the diffusion-scaled work-conserving action space $\hat{\mathcal{Z}}^n(\hat{x})$ takes the form

$$\hat{\mathcal{Z}}^n(\hat{x}) := \bigl\{\hat{z} : \sqrt{n}\hat{z} + n\bar{z}^n \in \mathcal{Z}^n\bigl(\sqrt{n}\hat{x} + n\bar{x}^n\bigr)\bigr\}\,, \qquad \hat{x} \in \mathcal{S}^n\,.$$

Recall that a scheduling policy is called stationary Markov if $Z^n(t) = z(X^n(t))$ for some function $z\colon \mathbb{Z}_+^m \to \mathbb{Z}_+^{\mathcal{G}}$, in which case we identify the policy with the function z. Under a stationary Markov policy, X^n is Markov with controlled generator

$$\mathcal{L}_z^n f(x) := \sum_{i\in\mathcal{I}} \Biggl(\lambda_i^n\bigl(f(x+e_i) - f(x)\bigr) + \sum_{j\in\mathcal{J}(i)} \mu_{ij}^n z_{ij}\bigl(f(x-e_i) - f(x)\bigr)\Biggr) \tag{2.13}$$

for $f \in \mathcal{C}(\mathbb{R}^m)$ and $x \in \mathbb{Z}^m_+$. Let $\ell^n = (\ell^n_1, \dots, \ell^n_m)^{\mathsf{T}}$ be defined by

$$\ell^n_i \;:=\; \frac{1}{\sqrt{n}}\biggl(\lambda^n_i - \sum_{j\in\mathcal{J}(i)} \mu^n_{ij}\xi^*_{ij}N^n_j\biggr)\,. \tag{2.14}$$

By (2.8), the assumptions on the parameters in (2.2) and (2.3), we have

$$\ell^n_i \;\xrightarrow[n\to\infty]{}\; \ell_i \;:=\; \hat\lambda_i - \sum_{j\in\mathcal{J}(i)} \hat\mu_{ij} z^*_{ij} - \sum_{j\in\mathcal{J}(i)} \mu_{ij}\xi^*_{ij}\hat\nu_j\,,$$

with z^* as in (2.5). Let $\ell := (\ell_1, \dots, \ell_m)^{\mathsf{T}}$. Note that ℓ^n_i and ℓ_i can be regarded as the deficit or surplus in the number of servers (of order $\mathcal{O}(\sqrt{n})$ allocated to class i in the diffusion scale. Note also that ℓ^n_i and ℓ_i appears as constants in the drift of the diffusion-scaled and diffusion limit processes; see (2.25) and (2.30)

We drop the dependence on n in the diffusion-scaled variables in order to simplify the notation. A work-conserving stationary Markov policy z, that is a map $z\colon \mathbb{Z}^m_+ \to \mathbb{Z}^{\mathcal{G}}_+$ such that $z(x) \in \mathcal{Z}^n(x)$ for all $x \in \mathbb{Z}^m_+$, gives rise to a policy $\hat z\colon \mathcal{S}^n \to \mathbb{R}^{\mathcal{G}}$, with $\hat z(\hat x) \in \hat{\mathcal{Z}}^n(\hat x)$ for all $\hat x \in \mathcal{S}^n$, via (2.11) (and vice-versa). Using (2.9), (2.13), and (2.14) and rearranging terms, the controlled generator of the corresponding diffusion-scaled process can be written as

$$\begin{aligned}\widehat{\mathcal{L}}^n_{\hat z} f(\hat x) \;=\; &\sum_{i\in\mathcal{I}} \frac{\lambda^n_i}{n}\, \frac{\mathfrak{d}f\bigl(\hat x; \frac{1}{\sqrt{n}}e_i\bigr) + \mathfrak{d}f\bigl(\hat x; -\frac{1}{\sqrt{n}}e_i\bigr)}{n^{-1}} \\ &- \sum_{i\in\mathcal{I}} b^n_i(\hat x, \hat z)\, \frac{\mathfrak{d}f\bigl(\hat x; -\frac{1}{\sqrt{n}}e_i\bigr)}{n^{-1/2}}\,, \quad \hat x \in \mathcal{S}^n\,,\ \hat z \in \hat{\mathcal{Z}}^n(\hat x)\,,\end{aligned} \tag{2.15}$$

where $\mathfrak{d}f$ is given by

$$\mathfrak{d}f(x; y) \;:=\; f(x+y) - f(x)\,, \quad x, y \in \mathbb{R}^m\,,$$

and the 'drift' $b^n = (b^n_1, \cdots, b^n_m)^{\mathsf{T}}$ is given by

$$b^n_i(\hat x, \hat z) \;:=\; \ell^n_i - \sum_{j\in\mathcal{J}(i)} \mu^n_{ij}\hat z_{ij}\,, \quad \hat z \in \hat{\mathcal{Z}}^n(\hat x)\,,\ i \in \mathcal{I}\,. \tag{2.16}$$

Abusing the notation for $\hat x \in \mathcal{S}^n$ and $\hat z \in \hat{\mathcal{Z}}^n(\hat x)$, we define (compare with (2.10))

$$\hat q^n_i(\hat x, \hat z) \;:=\; \hat x_i - \sum_{j\in\mathcal{J}(i)} \hat z_{ij}\,, \quad i \in \mathcal{I}\,, \qquad \hat y^n_j(\hat z) \;:=\; -\sum_{i\in\mathcal{I}(j)} \hat z_{ij}\,, \quad j \in \mathcal{J}\,, \tag{2.17}$$

and

$$\hat\vartheta^n(\hat x, \hat z) \;:=\; \langle e, \hat q^n(\hat x, \hat z)\rangle \wedge \langle e, \hat y^n(\hat z)\rangle\,. \tag{2.18}$$

Recall (2.9). The parameter $\hat\vartheta^n$ can therefore be regarded as the scaled minimum of the total number of customers in the queues and the total number of idle servers.

By (2.17), we have

$$\bigl\langle e, \hat q^n(\hat x, \hat z)\bigr\rangle \;=\; \hat\vartheta^n(\hat x, \hat z) + \langle e, \hat x\rangle^+\,, \quad \text{and} \quad \bigl\langle e, \hat y^n(\hat z)\bigr\rangle \;=\; \hat\vartheta^n(\hat x, \hat z) + \langle e, \hat x\rangle^- \tag{2.19}$$

for all $\hat x \in \mathcal{S}^n$ and $\hat z \in \hat{\mathcal{Z}}^n(\hat x)$. Define the $(m-1)$ and $(J-1)$ simplexes

$$\varDelta_c \;:=\; \{u \in \mathbb{R}^m : u \ge 0\,,\ \langle e, u\rangle = 1\}\,, \quad \text{and} \quad \varDelta_s \;:=\; \{u \in \mathbb{R}^J : u \ge 0\,,\ \langle e, u\rangle = 1\}\,, \tag{2.20}$$

and let $\varDelta := \varDelta_c \times \varDelta_s$. By (2.19), there exists $u = (u^c, u^s) \in \varDelta$ such that

$$\hat q^n(\hat x, \hat z) \;=\; \bigl(\hat\vartheta^n(\hat x, \hat z) + \langle e, \hat x\rangle^+\bigr)\, u^c\,, \quad \text{and} \quad \hat y^n(\hat z) \;=\; \bigl(\hat\vartheta^n(\hat x, \hat z) + \langle e, \hat x\rangle^-\bigr)\, u^s\,. \tag{2.21}$$

Let

$$\mathcal{D} \;:=\; \Bigl\{(\alpha, \beta) \in \mathbb{R}^m \times \mathbb{R}^J :\ \textstyle\sum_{i=1}^m \alpha_i = \sum_{j=1}^J \beta_j\Bigr\}\,.$$

As shown in [10, Proposition A.2], there exists a unique linear map $\Phi = [\Phi_{ij}]\colon \mathcal{D} \to \mathbb{R}^{\mathcal{G}}$ solving

$$\sum_{j\in\mathcal{J}(i)} \Phi_{ij}(\alpha,\beta) \,=\, \alpha_i \quad \forall\, i \in \mathcal{I}\,, \quad \text{and} \quad \sum_{i\in\mathcal{I}(j)} \Phi_{ij}(\alpha,\beta) \,=\, \beta_j \quad \forall\, j \in \mathcal{J}\,. \tag{2.22}$$

The solutions Φ_{ij} correspond to the resource allocations $\hat{z}_{ij}$ in the diffusion scale, see (2.23). Since $\bigl(\hat{x} - \hat{q}^n(\hat{x},\hat{z}), -\hat{y}^n(\hat{z})\bigr) \in \mathcal{D}$ by (2.7) and (2.17), using the linearity of the map Φ and (2.21) and (2.22), it follows that

$$\begin{aligned} \hat{z} \,&=\, \Phi\bigl(\hat{x} - \hat{q}^n(\hat{x},\hat{z}), -\hat{y}^n(\hat{z})\bigr) \\ &=\, \Phi\bigl(\hat{x} - \langle e,\hat{x}\rangle^+ u^c, -\langle e,\hat{x}\rangle^- u^s\bigr) - \hat{\vartheta}^n(\hat{x},\hat{z})\,\Phi(u^c,u^s)\,. \end{aligned} \tag{2.23}$$

We describe an important property of the linear map Φ which we need later. Consider the matrices $B_1^n \in \mathbb{R}^{m\times m}$ and $B_2^n \in \mathbb{R}^{m\times J}$ defined by

$$\sum_{j\in\mathcal{J}(i)} \mu_{ij}^n \Phi_{ij}(\alpha,\beta) \,=\, \bigl(B_1^n\alpha + B_2^n\beta\bigr)_i\,, \quad \forall\, i \in \mathcal{I}\,,\ \forall (\alpha,\beta)\in\mathcal{D}\,. \tag{2.24}$$

It is clear that for B_1^n to be a nonsingular matrix the basis used in the representation of the linear map Φ should be of the form $\mathfrak{D} = \bigl(\alpha, (\beta)_{-j}\bigr)$, $j \in \mathcal{J}$, where $(\beta)_{-j} = \{\beta_\ell\,,\ \ell \neq j\}$. Since Φ has a unique representation in terms of such a basis, and since B_i^n, $i = 1,2$, are determined uniquely from Φ by (2.24), abusing the terminology, we refer to such an $\mathfrak{D}$ as a basis for B_i^n, $i = 1,2$. In [5, Lemma 4.3], the following property is asserted: Given any $\hat{\imath} \in \mathcal{I}$, there exists an ordering of $\{\alpha_i\,,\ i \in \mathcal{I}\}$ with $\alpha_{\hat{\imath}}$ the last element, and $\hat{\jmath} \in \mathcal{J}$, such that the matrix B_1^n is lower diagonal with positive diagonal elements with respect to this ordered basis $\bigl(\alpha, (\beta)_{-\hat{\jmath}}\bigr)$. For more details, we refer the reader to [5, Section 4.1].

In view of (2.23) and (2.24), for any $\hat{z} \in \widetilde{\mathcal{Z}}^n(\hat{x})$ with $\hat{x} \in \mathcal{S}^n$, there exists $u = u(\hat{x},\hat{z}) \in \mathit{\Delta}$ such that the drift b^n in (2.16) takes the form

$$b^n(\hat{x},\hat{z}) \,=\, \ell^n - B_1^n\bigl(\hat{x} - \langle e,\hat{x}\rangle^+ u^c\bigr) + B_2^n u^s \langle e,\hat{x}\rangle^- + \hat{\vartheta}^n(\hat{x},\hat{z})\bigl(B_1^n u^c + B_2^n u^s\bigr)\,. \tag{2.25}$$

2.3. **Joint and system-wide work conservation.** We start with the following definition.

Definition 2.1. We say that an action $\hat{z} \in \hat{\mathcal{Z}}^n(\hat{x})$ is *jointly work conserving* (JWC), if $\hat{\vartheta}^n(\hat{x},\hat{z}) = 0$. Recall that a work conserving policy refers to an action in which a server is idle if and only if there is no customer waiting in the queue that this server can serve. A jointly work conserving action keeps all servers busy unless all queues are empty. Let

$$\hat{\vartheta}_*^n(\hat{x}) \,:=\, \min_{\hat{z}\in\hat{\mathcal{Z}}^n(\hat{x})} \hat{\vartheta}^n(\hat{x},\hat{z})\,, \quad \hat{x} \in \mathcal{S}^n\,,$$

and

$$\widetilde{\mathcal{Z}}^n(\hat{x}) \,:=\, \bigl\{\hat{z} \in \hat{\mathcal{Z}}^n(\hat{x}) : \hat{\vartheta}^n(\hat{x},\hat{z}) = \hat{\vartheta}_*^n(\hat{x})\bigr\}\,, \quad \hat{x} \in \mathcal{S}^n\,.$$

We refer to $\widetilde{\mathcal{Z}}^n(\hat{x})$ as the *system-wide work conserving* (SWC) action set at $\hat{x}$. A stationary Markov scheduling policy $\hat{z}$ is called SWC if $\hat{z}(\hat{x}) \in \widetilde{\mathcal{Z}}^n(\hat{x})$ for all $\hat{x} \in \mathcal{S}^n$. We let $\widetilde{\mathfrak{Z}}^n$ denote the class of all such policies. Since z and $\hat{z}$ are related by (2.11), abusing this terminology, we also refer to a Markov policy $z\colon \mathbb{Z}_+^m \to \mathbb{Z}_+^{\mathcal{G}}$ as SWC, if it satisfies

$$\frac{z(x) - n\bar{z}^n}{\sqrt{n}} \,\in\, \widetilde{\mathcal{Z}}^n\biggl(\frac{x - n\bar{x}^n}{\sqrt{n}}\biggr)\,,$$

and we write $z \in \widetilde{\mathfrak{Z}}^n$.

We recall [11, Lemma 3] which states that there exists $M_0 > 0$ such that the collection of sets $\breve{\mathfrak{X}}^n$ defined by

$$\breve{\mathfrak{X}}^n \,:=\, \bigl\{\hat{x} \in \mathcal{S}^n : \|\hat{x}\|_1 \le M_0\sqrt{n}\bigr\}\,, \tag{2.26}$$

has the following property. If $\hat{x} \in \breve{\mathfrak{X}}^n$, then for any pair $(\hat{q}, \hat{y})$ such that $\sqrt{n}\hat{q} \in \mathbb{Z}^m_+$, $\sqrt{n}\hat{y} \in \mathbb{Z}^J_+$, and satisfying

$$\langle e, \hat{q}\rangle \wedge \langle e, \hat{y}\rangle \,=\, 0\,, \quad \langle e, \hat{x} - \hat{q}\rangle \,=\, \langle e, -\hat{y}\rangle\,, \quad \text{and} \quad \hat{y}_j \le N^n_j\,, \quad j \in \mathcal{J}\,,$$

it holds that $\Phi(\hat{x} - \hat{q}, -\hat{y}) \in \hat{\mathcal{Z}}^n(\hat{x})$. It follows from this lemma and Definition 2.1 that if $\hat{x} \in \breve{\mathfrak{X}}^n$, then the actions in $\widetilde{\mathcal{Z}}^n(\hat{x})$ are JWC.

Remark 2.1. Using (2.19), we know that under the JWC condition

$$\bigl\langle e, \hat{q}^n(\hat{x}, \hat{z})\bigr\rangle \,=\, \langle e, \hat{x}\rangle^+\,, \quad \text{and} \quad \bigl\langle e, \hat{y}^n(\hat{z})\bigr\rangle \,=\, \langle e, \hat{x}\rangle^-\,.$$

Rewriting (2.21) under the JWC condition, we therefore have the following

$$u^c_i \,=\, \begin{cases} \frac{\hat{q}^n_i(\hat{x},\hat{z})}{\langle e, \hat{q}^n(\hat{x},\hat{z})\rangle}\,, & \text{if } \bigl\langle e, \hat{q}^n(\hat{x}, \hat{z})\bigr\rangle > 0\,,\\ e_1\,, & \text{otherwise} \end{cases}$$

and

$$u^s_j \,=\, \begin{cases} \frac{\hat{y}^n_j(\hat{z})}{\langle e, \hat{y}^n(\hat{z})\rangle}\,, & \text{if } \bigl\langle e, \hat{y}^n(\hat{z})\bigr\rangle > 0\,,\\ e_1\,, & \text{otherwise.} \end{cases}$$

Hence, one can see that the control u^c_i represents the proportion of the total queue length in the network at queue i, while u^s_j represents the proportion of the total number of idle servers in the network at pool j.

Recall from Definition 2.1 that a JWC action keeps all servers busy unless all queues are empty. It is clear that this cannot be always enforced in multiclass multi-pool networks over the entire state space. A SWC control enforces the complementarity between overall queue length and service.

2.4. **The diffusion limit.** The *diffusion approximation* or *diffusion limit* of the queueing model described above is an m-dimensional stochastic differential equation (SDE) of the form (see [10] and [11, Section 2.5])

$$\mathrm{d}X_t \,=\, b(X_t, U_t)\,\mathrm{d}t + \upsigma(X_t)\,\mathrm{d}W_t\,, \qquad X_0 = x \in \mathbb{R}^m\,. \tag{2.27}$$

Here, $\{W_t\}_{t\ge 0}$ is a standard m-dimensional Brownian motion, and the control U_t takes values in the set $\mathit{\Delta} = \mathit{\Delta}_c \times \mathit{\Delta}_s$ defined in (2.20). The drift b can be derived as follows. Recall $\mathbb{R}^{\mathcal{G}}$ in (2.1). For $u = (u^c, u^s) \in \mathit{\Delta}$, let $\widehat{\Phi}[u]\colon \mathbb{R}^m \to \mathbb{R}^{\mathcal{G}}$ be defined by

$$\widehat{\Phi}[u](x) \,\coloneqq\, \Phi\bigl(x - (e\cdot x)^+ u^c, -(e\cdot x)^- u^s\bigr)\,, \tag{2.28}$$

with Φ as defined in (2.22). Then the drift b takes the form

$$b_i(x, u) \,=\, \ell_i - \sum_{j\in\mathcal{J}(i)} \mu_{ij}\widehat{\Phi}_{ij}[u](x)\,. \tag{2.29}$$

By [5, Lemma 4.3], we also know that (2.29) can be expressed as

$$b(x, u) \,=\, \ell - B_1\bigl(x - \langle e, x\rangle^+ u^c\bigr) + B_2 u^s \langle e, x\rangle^-\,, \tag{2.30}$$

where $B_1 \in \mathbb{R}^{m\times m}$ is a lower diagonal matrix with positive diagonal elements, and $B_2 \in \mathbb{R}^{m\times J}$. Of course B_i in (2.30) and B^n_i in (2.25), $i = 1, 2$, have the same functional form with respect to $\{\mu_{ij}\}$ and $\{\mu^n_{ij}\}$, respectively.

The diffusion matrix $\upsigma \in \mathbb{R}^{m\times m}$ is constant, and

$$a \,\coloneqq\, \upsigma\upsigma^{\mathsf{T}} \,=\, \operatorname{diag}(2\lambda_1, \dots, 2\lambda_m)\,.$$

In addition, for $f \in \mathcal{C}^2(\mathbb{R}^m)$, we define

$$\mathcal{L}_u f(x) \,\coloneqq\, \frac{1}{2}\operatorname{trace}\bigl(a\nabla^2 f(x)\bigr) + \bigl\langle b(x, u), \nabla f(x)\bigr\rangle\,, \tag{2.31}$$

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

In the following, we describe the drift of the diffusion limit of the networks under consideration while showing some examples in Figure 1.

2.4.1. *Networks with a dominant server pool.* This network has one non-leaf server node, which, without loss of generality, we label as $j = 1$. As in Subsection 2.1, the customer nodes are denoted by $\mathcal{I} = \{1, 2, \dots, m\}$, and the server nodes by $\mathcal{J} = \{1, 2, \dots, J\}$. Recall that $\mathcal{J}(i)$ is the collection of sever nodes connected to customer i. Owing to the tree structure of the network, server $1 \in \mathcal{J}(i)$ for all $i \in \{1, 2, \dots, m\}$. Let $\mathcal{J}_1(i) \coloneqq \mathcal{J}(i) \setminus \{1\}$ for all $i \in \mathcal{I}$. Recall the form of the drift in (2.29). Using (2.22), it is simple to show that the matrix $\widehat{\Phi}_{ij}[u]$ for this network is given by

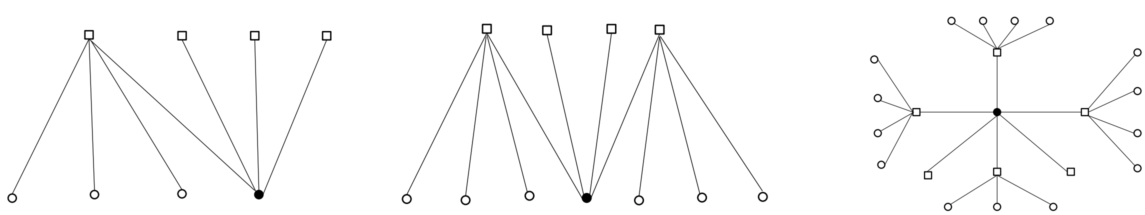
$$\widehat{\Phi}_{ij}[u](x) \;=\; \begin{cases} x_i - \langle e, x\rangle^+ u^c_i + \sum_{j\in\mathcal{J}_1(i)} \langle e, x\rangle^- \, u^s_j & \text{for } j = 1\,, \\ -\langle e, x\rangle^- \, u^s_j & \text{for } j \in \mathcal{J}_1(i)\,, \\ 0 & \text{otherwise.} \end{cases} \tag{2.32}$$

Using (2.32), the drift takes the following simple form:

$$b_i(x, u) \;=\; \ell_i - \mu_{i1}\bigl(x_i - u^c_i \langle e, x\rangle^+\bigr) + \sum_{j\in\mathcal{J}_1(i)} \mu_{i1}\bigl(\eta_{ij} - 1\bigr) u^s_j \langle e, x\rangle^-\,, \qquad i \in \mathcal{I}\,, \tag{2.33}$$

with $\eta_{ij} \coloneqq \frac{\mu_{ij}}{\mu_{i1}}$ for $j \in \mathcal{J}_1(i)$ and $i \in \mathcal{I}$. Note that $B_1 = \operatorname{diag}(\mu_{11}, \dots, \mu_{m1})$, and so $\ell = -\frac{\varrho}{m} B_1 e$, where ϱ is given by (2.35). We define

$$\bar\eta \;\coloneqq\; \max_{i\in\mathcal{I}}\; \max_{j\in\mathcal{J}_1(i)}\; \eta_{ij}\,, \qquad \text{and} \quad \underline{\eta} \;\coloneqq\; \min_{i\in\mathcal{I}}\; \min_{j\in\mathcal{J}_1(i)}\; \eta_{ij}\,.$$



(A) Generalized 'N' Network (B) Generalized 'M' Network (C) Network with a dominant server pool

FIGURE 1. Examples of multiclass multi-pool networks with a dominant pool (square–customer classes, circle–server pools, solid circle–the dominant server pool)

2.4.2. *Networks with class-dependent service rates.* We consider in this part arbitrary tree networks where the service rates are dictated by the customer type; namely $\mu_{ij} = \mu_i$ for all $(i, j) \in \mathcal{E}$. Recall the definition in (2.28). Using (2.22) and (2.29), the drift of this network takes the form

$$b_i(x, u) \;=\; \ell_i - \sum_{j\in\mathcal{J}(i)} \mu_{ij} \widehat{\Phi}_{ij}[u](x) \;=\; \ell_i - \mu_i\bigl(x_i - u^c_i \langle e, x\rangle^+\bigr)\,, \qquad \forall i \in \mathcal{I}\,. \tag{2.34}$$

Note then that $B_1 = \operatorname{diag}(\mu_1, \dots, \mu_m)$ and $B_2 = 0$.

We remark that both classes of networks have one common feature: the matrix B_1 is diagonal.

2.5. **A necessary and sufficient condition for uniform stability.** We define the spare capacity (or the safety staffing) for the n^{th} system (prelimit) and the diffusion limit by

$$\varrho_n := -\left\langle e(B_1^n)^{-1}, \ell^n \right\rangle, \quad \text{and} \quad \varrho := -\left\langle eB_1^{-1}, \ell \right\rangle, \tag{2.35}$$

respectively. Note, of course, that $\varrho_n \to \varrho$ as $n \to \infty$ by (2.2). Recall the expressions of ℓ^n and ℓ in (2.14). Recall that ℓ_i^n and ℓ_i can be regarded as the deficit or surplus in the number of servers (of order $\mathcal{O}(\sqrt{n})$ allocated to class i in the diffusion scale. Hence ϱ_n and ϱ can be regarded as the optimal reallocation of the capacity fluctuations (positive or negative) of order $\sqrt{n}$ when each server pool employs a square-root staffing rule.

We summarize the main results of the paper.

Theorem 2.1. *Consider a network with a dominant server pool, or with class-dependent service rates. Then the conditions $\varrho > 0$ and $\varrho_n > 0$ are necessary and sufficient for the uniform stability of the limiting diffusion and the diffusion-scaled queueing processes, respectively. More precisely:*

(i) *if $\varrho < 0$, the process $\{X_t\}_{t\geq 0}$ in (2.27) is transient under any stationary Markov control. In addition, if $\varrho = 0$, then $\{X_t\}_{t\geq 0}$ cannot be positive recurrent.*
(ii) *if $\varrho_n < 0$, the process $\{X_t^n\}_{t\geq 0}$ is transient under any stationary Markov scheduling policy. In addition, if $\varrho_n = 0$, then $\{X_t^n\}_{t\geq 0}$ cannot be positive recurrent.*
(iii) *if $\varrho > 0$, the processes $\{X_t\}_{t\geq 0}$ are uniformly exponentially ergodic over stationary Markov controls.*
(iv) *if $\varrho_n > 0$, the processes $\{X_t^n\}_{t\geq 0}$ are uniformly exponentially ergodic over SWC scheduling policies, and the invariant distributions have exponential tails.*

Parts (i) and (ii) of Theorem 2.1 follow from Propositions 3.1 and 3.2, respectively. Part (iii) follows from Theorem 4.2, and part (iv) from Theorem 5.1.

3. Two properties of the spare capacity and transience

In the first part of this section we prove the results in Theorem 2.1 (i) and (ii). It is important to note that for the models in Subsections 2.4.1 and 2.4.2 we have

$$1 + \langle e, B_1^{-1} B_2 u^s \rangle > 0\,. \tag{3.1}$$

Note that for networks with a dominant server pool, we have $1 = u_1^s + \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}_1(i)} u_j^s$. Hence

$$\begin{aligned}
1 + \langle e, B_1^{-1} B_2 u^s \rangle &= 1 + \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}_1(i)} (\eta_{ij} - 1) u_j^s \\
&= 1 - \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}_1(i)} u_j^s + \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}_1(i)} \eta_{ij} u_j^s \\
&= u_1^s + \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}_1(i)} \eta_{ij} u_j^s > 0.
\end{aligned}$$

Note also that for networks with a class dependent service rate we have established that $B_2 = 0$. Hence $1 + \langle e, B_1^{-1} B_2 u^s \rangle = 1 > 0$.

Proposition 3.1. *Suppose that $\varrho = -\langle e, B_1^{-1}\ell \rangle < 0$. Then the process $\{X_t\}_{t\geq 0}$ in (2.27) is transient under any stationary Markov control. In addition, if $\varrho = 0$, then $\{X_t\}_{t\geq 0}$ cannot be positive recurrent.*

Proof. Recall that the first and second order derivatives of the hyperbolic tangent function are

$$\tanh'(x) = \frac{1}{\cosh^2(x)}\,; \qquad \tanh''(x) = -2\frac{\tanh(x)}{\cosh^2(x)}$$

Let $H(x) := \tanh\bigl(\beta\langle e, B_1^{-1}x\rangle\bigr)$, with $\beta > 0$. Then

$$\mathrm{trace}\bigl(a\nabla^2 H(x))\bigr) \;=\; \beta^2 \tanh''\bigl(\beta\langle e, B_1^{-1}x\rangle\bigr)\bigl|\upsigma^{\mathsf{T}} B_1^{-1}e\bigr|^2\,.$$

We have

$$\begin{aligned}\mathcal{L}_u H(x) \;&=\; \frac{1}{2}\,\mathrm{trace}\bigl(a\nabla^2 H(x)\bigr) + \bigl\langle b(x,u), \nabla H(x)\bigr\rangle \\ &=\; -\beta^2\frac{\tanh\bigl(\beta\langle e, B_1^{-1}x\rangle\bigr)}{\cosh^2\bigl(\beta\langle e, B_1^{-1}x\rangle\bigr)}|\upsigma^{\mathsf{T}} B_1^{-1}e|^2 \\ &\qquad + \frac{\beta}{\cosh^2\bigl(\beta\langle e, B_1^{-1}x\rangle\bigr)}\biggl(\bigl\langle e, B_1^{-1}\ell\bigr\rangle + \langle e, x\rangle^-\Bigl(1 + \bigl\langle e, B_1^{-1}B_2 u^s\bigr\rangle\Bigr)\biggr)\,.\end{aligned} \tag{3.2}$$

Thus, for $0 < \beta < \langle e, B_1^{-1}\ell\rangle\,|\upsigma^{\mathsf{T}} B_1^{-1}e|^{-2}$, we obtain $\mathcal{L}_u H(x) > 0$ by (3.1). Therefore, $\{H\bigl(X_t\bigr)\}_{t\ge 0}$ is a bounded submartingale, so it converges almost surely. Since X is irreducible, it can be either recurrent or transient. If it is recurrent, then H should be constant a.e. in $\mathbb{R}^m$, which is not the case. Thus X is transient.

We now turn to the case where $\varrho = 0$. Suppose that the process $\{X(t)\}_{t\ge0}$ (under some stationary Markov control) has an invariant probability measure $\uppi(\mathrm{d}x)$. It is well known that $\uppi$ must have a positive density. Let $h_1(x)$ and $h_2(x)$ denote respectively the first and the second terms on the right hand side of (3.2). Applying Itô's formula to (3.2) with $X_0 = x$ as in (2.27), we obtain

$$\mathbb{E}^{\uppi}\bigl[H(X_{t\wedge\tau_r})\bigr] - H(x) \;=\; \sum_{i=1,2}\mathbb{E}^{\uppi}\biggl[\int_0^{t\wedge\tau_r} h_i(X_s)\mathrm{d}s\biggr]\,, \tag{3.3}$$

where τ_r denotes the first exit time from $\mathcal{B}_r$, $r > 0$. Note that $h_1(x)$ is bounded and $h_2(x)$ is non-negative. Thus using dominated and monotone convergence, we can take limits in (3.3) as $r \to \infty$ for the terms on the right side to obtain

$$\int_{\mathbb{R}^m} H(x)\uppi(\mathrm{d}x) - H(x) \;=\; t\sum_{i=1,2}\int_{\mathbb{R}^m} h_i(x)\uppi(\mathrm{d}x), \qquad t \ge 0\,.$$

Since $H(x)$ is bounded, we can divide both sides by t and β and take the limit as $t \to \infty$ to get

$$\int_{\mathbb{R}^m}\beta^{-1}h_1(x)\uppi(\mathrm{d}x) + \int_{\mathbb{R}^m}\beta^{-1}h_2(x)\uppi(\mathrm{d}x) \;=\; 0\,. \tag{3.4}$$

Since $\beta^{-1}h_1(x)$ tends to 0 uniformly in x as $\beta \searrow 0$, the first term on the left hand side of (3.4) vanishes as $\beta \searrow 0$. However, since $\beta^{-1}h_2(x)$ is bounded away from 0 on the open set $\{x \in \mathbb{R}^m \colon \langle e, x\rangle^- > 1\}$, this contradicts the fact that $\uppi(\mathrm{d}x)$ has full support. □

Remark 3.1. In the proof of Proposition 3.1, the function $H(x)$ is a bounded test function which was chosen so that $\mathcal{L}_u H(x) \ge 0$. Assume that X_t is recurrent. Since $H(X_t)$ converges a.s. (being a bounded submartingale), and since X_t 'visits every open neighborhood' in $\mathbb{R}^m$ with probability 1, it follows that H must be a constant function which is a contradiction.

The formalism behind the above argument is as follows: Let $\uptau$ be the first hitting time to the unit ball B centered at $x = 0$. If $\mathbb{E}_x$ denotes the expectation operator on the canonical space of the Markov process $\{X_t\}_{t\ge0}$ with initial condition $X_0 = x$, then by Dynkin's formula we obtain $\mathbb{E}_x[H(X_\uptau)] \ge H(x)$. If the process is recurrent, then of course $\mathbb{P}_x(\uptau < \infty) = 1$, which gives $\mathbb{E}_x[H(X_\uptau)] \le \sup_{y\in B} H(y)$. Thus $\sup_{y\in B} H(y) \ge H(x)$ for all $x \in \mathbb{R}^m$ which is not true. Moving the center of the the ball B to an arbitrary point z, and denoting it as $B(z)$, we similarly have

$$\sup_{y\in B(z)} H(y) \ge H(x) \qquad \forall x, z \in \mathbb{R}^m\,.$$

This implies that $H(x) = $ constant which is a contradiction and hence X is transient.

Proposition 3.2. *Suppose that $\varrho_n < 0$. Then the state process $\{X^n_t\}_{t\ge 0}$ of the n^{th} system is transient under any stationary Markov scheduling policy. In addition, if $\varrho^n = 0$, then $\{X^n_t\}_{t\ge 0}$ cannot be positive recurrent.*

Proof. The proof mimics that of Proposition 3.1. We apply the function H in that proof to the operator $\widehat{\mathcal{L}}^n_{\hat z}$ in (2.15), and use the identity

$$H\Bigl(x \pm \tfrac{1}{\sqrt{n}} e_i\Bigr) - H(x) \mp \frac{1}{\sqrt{n}}\partial_{x_i} H(x) \;=\; \frac{1}{n}\int_0^1 (1-t)\,\partial_{x_i x_i} H\Bigl(x \pm \tfrac{t}{\sqrt{n}} e_i\Bigr)\,\mathrm{d}t\,, \tag{3.5}$$

to express the first and second order incremental quotients, together with (2.25) which implies that

$$\bigl\langle b^n(\hat x,\hat z), \nabla H(\hat x)\bigr\rangle \;=\; \frac{\beta}{\cosh^2\bigl(\beta\langle e,(B^n_1)^{-1}\hat x\rangle\bigr)}\biggl(\bigl\langle e,(B^n_1)^{-1}\ell^n\bigr\rangle \\ + \bigl(\hat\vartheta^n(\hat x,\hat z) + \langle e,\hat x\rangle^-\bigr)\Bigl(1 + \bigl\langle e,(B^n_1)^{-1}B^n_2 u^s\bigr\rangle\Bigr)\biggr)\,.$$

The rest follows exactly as in the proof of Proposition 3.1. □

3.1. **Spare capacity and average idleness.** It is shown in [4,8] that if the diffusion limit of the 'V' network with no abandonment has a invariant distribution π under some stationary Markov control, then ϱ represents the 'average idleness' of the system, that is, $\varrho = \int_{\mathbb{R}^m}\langle e,x\rangle^-\pi(\mathrm{d}x)$. In calculating this average for multi-pool networks, idle servers are not weighted equally across different pools and the term $\bigl\langle e, B_1^{-1}B_2u^s(x)\bigr\rangle$ appears in the expression, see (3.6). It is important to note that only the control on the idleness allocations among server pools u^s appears in the identity, and the control component u^c does not.

Theorem 3.1. *Consider a network with a dominant server pool, or with class-dependent service rates, and suppose that $\varrho > 0$. Let π_u denote the invariant invariant probability measure corresponding to a stationary Markov control $u \in \mathfrak{U}_{\mathsf{sm}}$, whose existence follows from Theorem 2.1 (iii). Then*

$$\varrho \;=\; \int_{\mathbb{R}^m}\Bigl(1 + \bigl\langle e, B_1^{-1}B_2u^s(x)\bigr\rangle\Bigr)\langle e,x\rangle^-\,\pi_u(\mathrm{d}x). \tag{3.6}$$

Proof. The proof is similar to that of [8, Corollary 5.1], but more involved for the multiclass multi-pool networks. We first recall some definitions and notations. Let $\chi_r(t)$, $\breve\chi_r(t)$, $r>1$ be smooth, concave and convex functions, respectively, defined by

$$\chi_r(t) \;=\; \begin{cases} t\,, & t \le r-1\,,\\ r-\frac{1}{2}\,, & t\ge r\,,\end{cases} \qquad\text{and}\quad \breve\chi_r(t) \;=\; \begin{cases} t\,, & t\ge 1-r\,,\\ \frac{1}{2}-r\,, & t\le -r\,.\end{cases}$$

Let $g_r(x) = \breve\chi_r\bigl(\langle e, B_1^{-1}x\rangle\bigr)$, and $f_r(x) = \chi_r\bigl(g_r(x)\bigr)$. A straightforward calculation shows that

$$\begin{aligned}\langle b(x,u), \nabla f_r(x)\rangle \;&=\; h_1(x) + h_2(x)\,,\\ \frac{1}{2}\,\mathrm{trace}\bigl(a(x)\nabla^2 f_r(x)\bigr) \;&=\; h_3(x) + h_4(x)\,,\end{aligned}$$

where

$$\begin{aligned} h_1(x) \;&\coloneqq\; -\varrho\chi_r'\bigl(f(x)\bigr)\,\breve\chi_r'\bigl(\langle e,B_1^{-1}x\rangle\bigr)\,,\\ h_2(x) \;&\coloneqq\; \bigl[1+\bigl\langle e,B_1^{-1}B_2u^s\bigr\rangle\bigr]\chi_r'\bigl(f(x)\bigr)\,\breve\chi_r'\bigl(\langle e,B_1^{-1}x\rangle\bigr)\langle e,x\rangle^-\,,\\ h_3(x) \;&\coloneqq\; \frac{1}{2}\chi_r''\bigl(f(x)\bigr)\,\bigl(\breve\chi_r'\bigl(\langle e,B_1^{-1}x\rangle\bigr)\bigr)^2\bigl|\sigma^{\mathsf T}B_1^{-1}e\bigr|^2\,,\\ h_4(x) \;&\coloneqq\; \frac{1}{2}\chi_r'\bigl(f(x)\bigr)\breve\chi_r''\bigl(\langle e,B_1^{-1}x\rangle\bigr)\bigl|\sigma^{\mathsf T}B_1^{-1}e\bigr|^2\,.\end{aligned}$$

We note that $g_r(x)$ is positive and bounded below away from 0, and $f_r(x)$ is smooth, bounded, and has bounded derivatives. Also note that h_i, $i = 1, 2, 3$, are bounded, and h_2 is nonnegative. Therefore, if $\{X(t)\}_{t\ge 0}$ is positive recurrent with an invariant probability measure $\pi_u(\mathrm{d}x)$, a straightforward application of Itô's formula shows that $\pi_u\bigl(\mathcal{L}_u f_r\bigr) = 0$. Therefore, we obtain

$$\pi_u(-h_1) \;=\; \pi_u(h_2) + \pi_u(h_3) + \pi_u(h_4). \tag{3.7}$$

By the definition of χ_r and $\breve{\chi}_r$, it is straightforward to verify that

$$\lim_{r\to\infty} \pi_u(h_3) \;=\; \lim_{r\to\infty} \pi_u(h_4) \;=\; 0\,. \tag{3.8}$$

In addition, using dominated convergence theorem

$$\begin{aligned} \lim_{r\to\infty} \pi_u(h_1) \;&=\; -\varrho\,,\\ \lim_{r\to\infty} \pi_u(h_2) \;&=\; \int_{\mathbb{R}^m} \bigl(1 + \bigl\langle e, B_1^{-1} B_2 u^s\bigr\rangle\bigr) \langle e, x\rangle^-\, \pi_u(\mathrm{d}x)\,. \end{aligned} \tag{3.9}$$

Combining (3.7)–(3.9), we obtain (3.6). □

Remark 3.2. Using Theorem 3.1 and the drift in Subsections 2.4.1 and 2.4.2, it is easy to verify that in the case of a network with a dominant server pool we have

$$\varrho = \int_{\mathbb{R}^m} \Biggl[1 + \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}_1(i)} \bigl(\tfrac{\mu_{ij}}{\mu_{i1}} - 1\bigr) u^s_j\Biggr] \langle e, x\rangle^- \pi_u(\mathrm{d}x)\,,$$

whereas in the case of a network with class-dependent service rates, (3.6) takes the form

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

4. Uniform exponential ergodicity of the diffusion limit

In this section we show that if $\varrho > 0$ then the diffusion limit of a network with a dominant server pool, or with class-dependent service rates, is uniformly exponentially ergodic and the invariant distributions have exponential tails.

We start by reviewing the notion of *uniform exponential ergodicity* for a controlled diffusion. We do this under fairly general hypotheses. Consider a controlled diffusion process $X = \{X_t,\, t \ge 0\}$ which takes values in the m-dimensional Euclidean space $\mathbb{R}^m$, and is governed by the Itô equation

$$\mathrm{d}X_t \;=\; b\bigl(X_t, v(X_t)\bigr)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t\,. \tag{4.1}$$

All random processes in (4.1) live in a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The process W is a m-dimensional standard Wiener process independent of the initial condition X_0. The function v maps $\mathbb{R}^m$ to a compact, metrizable set $\mathbb{U}$ and is Borel measurable. The collection of such functions comprising of the set of stationary Markov controls is denoted by $\mathfrak{U}_{\mathsf{sm}}$.

The parameters of the equation (4.1) satisfy the following:

(1) *Local Lipschitz continuity:* The functions $b\colon \mathbb{R}^m \times \mathbb{U} \to \mathbb{R}^m$ and $\sigma\colon \mathbb{R}^m \to \mathbb{R}^{m\times m}$ are continuous, and satisfy

$$|b(x,u) - b(y,u)| + \|\sigma(x) - \sigma(y)\| \;\le\; C_R\,|x - y| \qquad \forall\, x, y \in B_R\,,\ \forall\, u \in \mathbb{U}\,.$$

for some constant $C_R > 0$ depending on $R > 0$.

(2) *Affine growth condition:* For some $C_0 > 0$, we have

$$\sup_{u\in\mathbb{U}}\, \langle b(x,u), x\rangle^+ + \|\sigma(x)\|^2 \;\le\; C_0\bigl(1 + |x|^2\bigr) \qquad \forall\, x \in \mathbb{R}^m\,.$$

(3) *Nondegeneracy:* For each $R > 0$, it holds that

$$\sum_{i,j=1}^{m} a^{ij}(x)\xi_i\xi_j \;\ge\; C_R^{-1}|\xi|^2 \qquad \forall\, x \in B_R\,,$$

and for all $\xi = (\xi_1, \dots, \xi_m)^{\mathsf{T}} \in \mathbb{R}^m$, where $a = \frac{1}{2}\upsigma\upsigma^{\mathsf{T}}$.

It is well known that, under hypotheses (1)–(2), (4.1) has a unique strong solution which is also a strong Markov process for any $v \in \mathfrak{U}_{\mathsf{sm}}$ [22]. We let $\mathbb{E}_x^v$ denote the expectation operator on the canonical space of the process controlled by v, with initial condition $X_0 = x$. Let $\uptau(A)$ denote the first exit time from the set $A \in \mathbb{R}^m$.

We say that the process $\{X_t\}_{t\ge0}$ is *uniformly exponentially ergodic* if for some ball $\mathcal{B}_\circ$ there exist $\delta_\circ > 0$ and $x_\circ \in \bar{\mathcal{B}}_\circ^{\mathsf{c}}$ such that $\sup_{v\in\mathfrak{U}_{\mathsf{sm}}} \mathbb{E}_{x_\circ}^v[\mathrm{e}^{\delta_\circ\,\uptau(\mathcal{B}_\circ^{\mathsf{c}})}] < \infty$.

We let $\widehat{\mathcal{A}}$ denote the operator

$$\widehat{\mathcal{A}}\phi(x) \;\coloneqq\; \frac{1}{2}\,\mathrm{trace}\bigl(a(x)\nabla^2\phi(x)\bigr) + \max_{u\in\mathbb{U}}\,\bigl\langle b(x,u), \nabla\phi(x)\bigr\rangle\,, \qquad x \in \mathbb{R}^m\,,$$

for $\phi \in \mathcal{C}^2(\mathbb{R}^m)$. For a locally bounded, Borel measurable function $f\colon \mathbb{R}^m \to \mathbb{R}$, which is bounded from below in $\mathbb{R}^m$, i.e., $\inf_{\mathbb{R}^m} f > -\infty$, we define the generalized principal eigenvalue of $\widehat{\mathcal{A}} + f$ by

$$\Lambda(f) \;\coloneqq\; \inf\Bigl\{\lambda \in \mathbb{R} : \exists\,\varphi \in \mathcal{W}^{2,m}_{\mathrm{loc}}(\mathbb{R}^m),\ \varphi > 0,\ \widehat{\mathcal{A}}\varphi + (f-\lambda)\varphi \le 0 \text{ a.e. in } \mathbb{R}^m\Bigr\}\,,$$

where $\mathcal{W}^{2,m}_{\mathrm{loc}}$ is a local Sobolev space. We have the following equivalent characterizations of uniform exponential ergodicity. This is a straightforward extension of [8, Theorem 3.1] for controlled diffusions, and is stated without proof. Recall that a map $f\colon \mathbb{R}^m \to \mathbb{R}$ is called *coercive*, or *inf-compact*, if $\inf_{\mathcal{B}_r^{\mathsf{c}}} f \to \infty$ as $r \to \infty$.

Theorem 4.1. *The following are equivalent.*

(*a*) *For some ball* $\mathcal{B}_\circ$ *there exists* $\delta_\circ > 0$ *and* $x_\circ \in \bar{\mathcal{B}}_\circ^{\mathsf{c}}$ *such that* $\sup_{v\in\mathfrak{U}_{\mathsf{sm}}} \mathbb{E}_{x_\circ}[\mathrm{e}^{\delta_\circ\,\uptau(\mathcal{B}_\circ^{\mathsf{c}})}] < \infty$.

(*b*) *For every ball* $\mathcal{B}$ *there exists* $\delta > 0$ *such that* $\sup_{v\in\mathfrak{U}_{\mathsf{sm}}} \mathbb{E}_x^v[\mathrm{e}^{\delta\,\uptau(\mathcal{B}^{\mathsf{c}})}] < \infty$ *for all* $x \in \mathcal{B}^{\mathsf{c}}$.

(*c*) *For every ball* $\mathcal{B}$, *there exists a coercive function* $\mathcal{V} \in \mathcal{W}^{2,p}_{\mathrm{loc}}(\mathbb{R}^m)$, $p > d$, *with* $\inf_{\mathbb{R}^m} \mathcal{V} \ge 1$, *and positive constants* κ_0 *and* δ *such that*

$$\widehat{\mathcal{A}}\,\mathcal{V}(x) \;\le\; \kappa_0\,\mathbb{1}_{\mathcal{B}}(x) - \delta\mathcal{V}(x) \qquad \forall\, x \in \mathbb{R}^m\,. \tag{4.2}$$

(*d*) *Equation (4.1) is recurrent, and* $\Lambda(\mathbb{1}_{\mathcal{B}^{\mathsf{c}}}) < 1$ *for every ball* $\mathcal{B}$.

Remark 4.1. Recall (1.2). Let $P_t^v(x, \mathrm{d}y)$ denote the transition probability of $\{X_t\}_{t\ge0}$ in (4.1) under a control $v \in \mathfrak{U}_{\mathsf{sm}}$. It is well known that (4.2) implies that there exist constants γ and C_γ which do not depend on the control v chosen, such that

$$\bigl\|P_t^v(x,\cdot\,) - \uppi_v(\cdot)\bigr\|_{\mathcal{V}} \;\le\; C_\gamma\mathcal{V}(x)\,\mathrm{e}^{-\gamma t}\,, \qquad \forall\, x \in \mathbb{R}^m\,,\ \forall\, t \ge 0\,, \tag{4.3}$$

where $\uppi_v$ denotes the invariant probability measure of $\{X_t\}_{t\ge0}$ under the control v. In addition, for any control $v \in \mathfrak{U}_{\mathsf{sm}}$ we have

$$\int_{\mathbb{R}^m} \mathcal{V}(x)\,\uppi_v(\mathrm{d}x) \;\le\; \frac{\kappa_0}{\delta}\,. \tag{4.4}$$

In particular for $\mathcal{V}(x)$ being an exponential function, the moment generating function of $\{X_t\}_{t\ge0}$ is finite.

4.1. **A class of intrinsic Lyapunov functions for the queueing network model.** As seen in Theorem 4.1, uniform exponential ergodicity is equivalent to the Foster–Lyapunov inequality in (4.2). In establishing this property for the diffusion limit of stochastic networks, a proper choice of a Lyapunov function is of tantamount importance. We first describe an intrinsic class of such functions.

We fix a 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

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

For $\varepsilon > 0$ we define

$$\psi_\varepsilon(t) \;\coloneqq\; \psi(\varepsilon t)\,,$$

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

Suppose that $B_1 = \operatorname{diag}(\tilde\mu_1, \dots, \tilde\mu_m)$. Using the function ψ_ε introduced above, we let

$$\Psi(x) \;\coloneqq\; \sum_{i\in\mathcal{I}} \frac{\psi_\varepsilon(x_i)}{\tilde\mu_i}\,, \tag{4.5}$$

with

$$\varepsilon \;\coloneqq\; \frac{\varrho}{3m}\left(\sum_{i\in\mathcal{I}} \frac{\lambda_i(3\tilde\mu_i + 2)}{\tilde\mu_i^2}\right)^{-1}. \tag{4.6}$$

We also define

$$V_1(x) \;\coloneqq\; \exp\bigl(\theta\Psi(-x)\bigr)\,, \quad V_2(x) \;\coloneqq\; \exp\bigl(\Psi(x)\bigr)\,, \quad \text{and} \;\; V(x) \;\coloneqq\; V_1(x) + V_2(x)\,, \tag{4.7}$$

with θ a positive constant.

As a result of fixing the value of ε in (4.6), Ψ depends only on the parameter θ. This simplifies the statements of the results in the rest of the paper.

We review some useful properties of the function ψ_ε. Note that for $\varepsilon > 0$ we have

$$\psi_\varepsilon'(t)t \;=\; \begin{cases} 0 & \varepsilon t \le -1\,, \\ \varepsilon t(\varepsilon t+1)^2(-2\varepsilon t+1) & \varepsilon t \in [-1,0]\,, \\ \varepsilon t & \varepsilon t \ge 0\,. \end{cases}$$

The minimum of $\psi_\varepsilon'(t)t$ when $\varepsilon t \in [-1,0]$ is $-\frac{3(39+55\sqrt{33})}{4096} \ge -\frac{1}{2}$ for $\varepsilon t = -\frac{1+\sqrt{33}}{16}$. Therefore

$$\begin{aligned} \sum_{i\in\mathcal{I}} \psi_\varepsilon'(x_i)x_i &= \sum_{i\in\mathcal{I}} \psi_\varepsilon'(x_i)x_i \mathbb{1}_{\{\varepsilon x_i \ge 0\}} + \sum_{i\in\mathcal{I}} \psi_\varepsilon'(x_i)x_i \mathbb{1}_{\{\varepsilon x_i \in [-1,0]\}} \\ &\ge \varepsilon \sum_{i\in\mathcal{I}} x_i \mathbb{1}_{\{x_i \ge 0\}} - \frac{1}{2}\sum_{i\in\mathcal{I}} \mathbb{1}_{\{\varepsilon x_i \in [-1,0]\}} \\ &\ge \varepsilon\|x^+\|_1 - \frac{m}{2}\,, \end{aligned} \tag{4.8}$$

and similarly $-\sum_{i\in\mathcal{I}} \psi_\varepsilon'(-x_i)x_i \;\ge\; \varepsilon\|x^-\|_1 - \frac{m}{2}$, where $\mathcal{I} \coloneqq \{1,2,\dots,m\}$. Note also that

$$-\sum_{i\in\mathcal{I}} \psi_\varepsilon'(-x_i)x_i \;\le\; \varepsilon\langle e, x\rangle \;\le\; \sum_{i\in\mathcal{I}} \psi_\varepsilon'(x_i)x_i\,.$$

The function V in (4.7), scaled by the parameter θ which are suppressed in the notation, is our choice of a Lyapunov function when B_1 is a diagonal matrix. It is constructed in an intricate manner in order to capture both the total workload (using $\Psi(x)$) on the positive half-space and the idleness (using $\Psi(-x)$) on the negative half-space. As one cannot simply use x^+ or x^-, we must construct mollified smooth functions for them. In addition, one must recognize that the effects of the workload and idleness on the system are not the same (not symmetric), so we also introduce a parameter θ in the definition of $V_1(\cdot)$. The reader will notice the similarities in (4.7) and [4, Definition 2.2]. However the function V used in this paper is the sum of the two exponential functions V_1 and V_2, whereas their product is used in [4, Lemma 2.1]. As will be seen later, in the case of multiclass multi-pool networks, the analysis is considerably more complex. In Subsection 4.3.1 for example, the proofs required $V_1 \ge V_2^2$ on a subset of the state space while requiring the opposite inequality on its complement in order to establish the Foster-Lyapunov inequality in (4.15). See the discussions following (4.16) and (4.22). This is mainly the reason behind using the sum instead of the product when defining the function V.

In the following subsections, we establish the uniform exponential ergodicity of the networks under consideration. To help with the exposition, we study the 'N' network in detail and then proceed to the more general networks with a dominant server pool, and networks with class-dependent service rates.

In establishing the desired drift inequalities, we often partition the space appropriately, and focus on the subsets of the partition. The following cones appear quite often in the analysis.

For $\delta \in [0,1]$, we define the cones

$$\begin{aligned} \mathcal{K}_\delta^+ &:= \bigl\{x \in \mathbb{R}^m : \langle e, x\rangle \ge \delta \|x\|_1\bigr\}\,, \\ \mathcal{K}_\delta^- &:= \bigl\{x \in \mathbb{R}^m : \langle e, x\rangle \le -\delta \|x\|_1\bigr\}. \end{aligned} \tag{4.9}$$

It is clear that $\mathcal{K}_0^+$ ($\mathcal{K}_0^-$) corresponds to the nonnegative (nonpositive) canonical half-space, and $\mathcal{K}_1^+$ ($\mathcal{K}_1^-$) is the nonnegative (nonpositive) closed orthant.

The following identities are very useful.

$$\langle e, x^+\rangle \;=\; \frac{1\pm\delta}{2}\,\|x\|_1\,, \qquad \langle e, x^-\rangle \;=\; \frac{1\mp\delta}{2}\,\|x\|_1 \qquad \text{for } \; x \in \partial\mathcal{K}_\delta^\pm\,,\; \delta\in[0,1]\,. \tag{4.10}$$

In addition, it is straightforward to show that

$$\sum_{i\in\mathcal{I}} \psi_\varepsilon(x_i) \;\le\; \sum_{i\in\mathcal{I}} \psi_\varepsilon(-x_i) \quad \text{if } \; x\in\mathcal{K}_0^-\,. \tag{4.11}$$

Also, the following inequality is true in general for any $\mathcal{I}' \subset \mathcal{I}$.

$$\sum_{i\in\mathcal{I}'} \psi'_\varepsilon(x_i) x_i - \varepsilon \sum_{i\in\mathcal{I}'} x_i \;=\; \sum_{x_i<0\,,\, i\in\mathcal{I}'} \bigl(\psi'_\varepsilon(x_i) - \varepsilon\bigr) x_i \;\ge\; 0\,. \tag{4.12}$$

Remark 4.2. There is an important scaling of the drift which we employ. Note that if we let $\zeta = \frac{\varrho}{m} e + B_1^{-1}\ell$, with ϱ as in (2.35), then a mere translation of the origin of the form $\tilde{X}_t = X_t + \zeta$ results in a diffusion of with the same drift as (2.29), except that the vector ℓ gets replaced by $\ell = -\frac{\varrho}{m} B_1 e$. Therefore, we may assume without any loss of generality that the drift in (2.30) takes the form

$$b(x,u) \;=\; -\frac{\varrho}{m} B_1 e - B_1\bigl(x - \langle e, x\rangle^+ u^c\bigr) + B_2 u^s \langle e, x\rangle^-. \tag{4.13}$$

4.2. **The Foster–Lyapunov inequality.** Recall the definition of the operator $\mathcal{L}_u$ in (2.31). We start with the following simple assertion.

Lemma 4.1. *Let V be the function in (4.7) with ε as in (4.6), and $\theta \ge \theta_0$ where θ_0 is a constant. Suppose that for any $\delta \in (0,1)$ there exist positive constants c_0 and c_1 such that the drift b in (4.13) satisfies*

$$\langle b(x,u), \nabla V(x)\rangle \;\le\; c_0 - \varepsilon c_1 \|x\|_1 V(x) \qquad \forall\,(x,u) \in \bigl(\mathcal{K}_\delta^+\bigr)^{\mathsf{c}} \times \mathit{\Delta}\,, \tag{4.14a}$$

$$\langle b(x,u), \nabla V_2(x)\rangle \;\le\; -\frac{\varrho\varepsilon}{m} V_2(x) \qquad \forall\,(x,u) \in \mathcal{K}_\delta^+ \times \mathit{\Delta}\,. \tag{4.14b}$$

Then, there exists a constant C_0 such that

$$\mathcal{L}_u V\bigl(x\bigr) \;\le\; C_0 - \frac{\varrho\varepsilon}{3m} V(x) \qquad\qquad \forall (x,u)\in\mathbb{R}^m\times\mathit{\Delta}\,. \tag{4.15}$$

Proof. A straightforward calculation, using the fact that $\psi''_\varepsilon(t)\le\frac{3}{2}\varepsilon^2$, shows that

$$\frac{1}{2}\,\mathrm{trace}\bigl(a\nabla^2 V_2(x)\bigr) \;\le\; \varepsilon^2\sum_{i\in\mathcal{I}}\frac{\lambda_i(3\mu_i+2)}{2\mu_i^2}\,V_2(x) \qquad \forall\, x\in\mathbb{R}^m\,.$$

Therefore, the choice of ε in (4.6) implies that $\frac{1}{2}\,\mathrm{trace}\bigl(a\nabla^2V_2(x)\bigr)\le\frac{\varrho\varepsilon}{4m}V_2$ for all $x\in\mathbb{R}^2$, and thus

$$\mathcal{L}_u V_2(x) \;\le\; -\frac{3\varrho\varepsilon}{4m}V_2(x) \qquad \forall\,(x,u)\in\mathcal{K}_\delta^+\times\mathit{\Delta} \tag{4.16}$$

by (4.14b). Since $|x^+|\ge\frac{1+\delta}{1-\delta}|x^-|$ for all $x\in\mathcal{K}_\delta^+$, we may select δ sufficiently close to 1 such that $V_2\ge V_1^2$ on $\mathcal{K}_\delta^+\cap\mathcal{K}_r^{\mathsf{c}}$ for some $r>0$. Since V_2 has exponential growth in $\|x\|_1$ on $\mathcal{K}_\delta^+$ and $\mathcal{L}_uV_1(x)\;\le\;C(1+|x|_1)V_1(x)$ on $\mathcal{K}_\delta^+\times\mathit{\Delta}$, it then follows that (4.15) holds on $\mathcal{K}_\delta^+\times\mathit{\Delta}$ by (4.16). It is also clear that (4.15) also holds on $\bigl(\mathcal{K}_\delta^+\bigr)^{\mathsf{c}}\times\mathit{\Delta}$ by (4.14a). This completes the proof. □

We now have the following result.

Theorem 4.2. *Assume that $\varrho>0$. Let V be the function in (4.7) with ε as in (4.6), and $\theta\ge\theta_0$ where θ_0 is a constant. Then the diffusion limit of any network with a dominant server pool or with class-dependent service rates is uniformly exponentially ergodic and the invariant distributions have exponential tails. In particular, there exists C_0 such that*

$$\mathcal{L}_u V\bigl(x\bigr) \;\le\; C_0 - \frac{\varrho\varepsilon}{3m} V(x) \qquad\qquad \forall (x,u)\in\mathbb{R}^m\times\mathit{\Delta}\,. \tag{4.17}$$

In particular,

$$\bigl\|P_t^v(x,\cdot\,)-\pi_v(\cdot)\bigr\|_V \;\le\; C_\gamma V(x)\,\mathrm{e}^{-\gamma t}\,, \qquad \forall\,x\in\mathbb{R}^m\,,\ \forall\,t\ge0\,, \tag{4.18}$$

where $P_t^v(x,\mathrm{d}y)$ denote the transition probability of $\{X_t\}_{t\ge0}$ under $v=(u^c,u^s)$ and π_v denotes the invariant probability measure of $\{X_t\}_{t\ge0}$ under the control v.

Proof. In Lemmas 4.3 and 4.4 in the section which follows, we establish (4.14a) and (4.14b) for these networks. Thus the proof of (4.17) follows directly from Lemma 4.1. □

4.3. **Three technical lemmas.** In this section, we prove (4.14a) and (4.14b) for the networks under consideration which implies Theorem 4.2. Even though the 'N' network is a special case of the networks with a dominant server pool, we first establish the result for this network in Lemma 4.2 as understanding the results in $\mathbb{R}^2$ will definitely help the reader in understanding the equations in Lemmas 4.3 and 4.4.

4.3.1. *The case of the 'N' network.* Here, $m=2$, and the matrices B_i, $i=1,2$, in (2.30) are given by

$$B_1 \;=\; \begin{pmatrix}\mu_{11} & 0\\ 0 & \mu_{21}\end{pmatrix}\,, \qquad \text{and}\;\; B_2 \;=\; \begin{pmatrix}0 & \mu_{12}-\mu_{11}\\ 0 & 0\end{pmatrix}.$$

Thus, using (4.13), the drift $b\colon\mathbb{R}^2\to\mathbb{R}^2$ for the 'N' network is given by

$$b(x,u) \;=\; -\frac{\varrho}{2}\begin{pmatrix}\mu_{11}\\ \mu_{21}\end{pmatrix} - \begin{pmatrix}\mu_{11} & 0\\ 0 & \mu_{21}\end{pmatrix}\bigl(x-\langle e,x\rangle^+u^c\bigr) + \begin{pmatrix}(\mu_{12}-\mu_{11})u_2^s\\ 0\end{pmatrix}\langle e,x\rangle^-\,. \tag{4.19}$$

Note that for the 'N' network, we have $\Psi(x)=\frac{\psi_\varepsilon(x_1)}{\mu_{11}}+\frac{\psi_\varepsilon(x_2)}{\mu_{21}}$ by (4.5). Recall the definition of the cube K_r in (1.1). We have the following lemma that verifies the drift inequalities (4.14a) and (4.14b) for the 'N' network.

Lemma 4.2. *Consider an 'N' network satisfying $\varrho > 0$. Let $\delta \in (0,1)$, $\theta \ge \theta_0 := 2(\eta \vee \eta^{-1})$, with $\eta := \frac{\mu_{12}}{\mu_{11}}$, and $V(x)$ be as in (4.7). Then, (4.14a) and (4.14b) hold with $m = 2$.*

Proof. To simplify the notation we define

$$F_i(x,u) := \frac{1}{V_i(x)} \bigl\langle b(x,u), \nabla V_i(x) \bigr\rangle\,, \quad i = 1,2\,. \tag{4.20}$$

We use (4.19), and apply (4.8) and the inequalities $\frac{\varrho}{2}\sum_{i\in\mathcal{I}} \psi'_\varepsilon \le \varrho\varepsilon$, and

$$\begin{aligned}\psi'_\varepsilon(-x_1)(\eta-1)u^s_2\langle e,x\rangle &\le -\varepsilon(1-\eta)^+\langle e,x\rangle\\ &\le \varepsilon(1-\eta)^+\|x^-\|_1 \qquad \forall\,(x,u)\in\mathcal{K}_0^-\times\Delta\,.\end{aligned}$$

to obtain

$$\begin{aligned}\frac{1}{\theta}\,F_1(x,u) &= \frac{\varrho}{2}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(-x_i) + \sum_{i\in\mathcal{I}}\psi'_\varepsilon(-x_i)x_i - \psi'_\varepsilon(-x_1)(\eta-1)u^s_2\langle e,x\rangle^-\\ &\le 1+\varrho\varepsilon-\varepsilon\|x^-\|_1+\varepsilon(1-\eta)^+\|x^-\|_1\\ &\le (1+\varrho\varepsilon)-\varepsilon(\eta\wedge 1)\|x^-\|_1\\ &\le (1+\varrho\varepsilon)-\frac{\varepsilon}{2}(\eta\wedge 1)\|x\|_1 \qquad \forall\,(x,u)\in\mathcal{K}_0^-\times\Delta\,.\end{aligned} \tag{4.21}$$

Similarly, we have

$$\begin{aligned}F_2(x,u) &= -\frac{\varrho}{2}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i) - \sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)x_i + \psi'_\varepsilon(x_1)(\eta-1)u^s_2\langle e,x\rangle^-\\ &\le \varepsilon\bigl(1+(\eta-1)^+\bigr)\|x\|_1\\ &\le \varepsilon(\eta\vee 1)\|x\|_1 \qquad \forall\,(x,u)\in\mathcal{K}_0^-\times\Delta\,.\end{aligned} \tag{4.22}$$

Note that, due to (4.11) and the choice of θ, we have $V_1 \ge V_2^2$ on $\mathcal{K}_0^-$. Thus, since V_1 has exponential growth in $\|x\|_1$ on $\mathcal{K}_0^-$, combining (4.21) and (4.22) and choosing an appropriate cube K_r, we obtain

$$\langle b(x,u), \nabla V(x)\rangle \le \Bigl(\theta(1+\varrho\varepsilon) - \frac{\varepsilon}{4}(\eta\wedge 1)\|x\|_1\Bigr)V(x) \qquad \forall\,(x,u)\in(\mathcal{K}_0^-\setminus K_r)\times\Delta\,. \tag{4.23}$$

We continue with estimates on $\mathcal{K}_0^+$. A straightforward calculation shows that

$$\left.\begin{aligned}\frac{1}{\theta}\,F_1(x,u) &= \frac{\varrho}{2}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(-x_i) + \sum_{i\in\mathcal{I}}\psi'_\varepsilon(-x_i)x_i - \sum_{i\in\mathcal{I}}\psi'_\varepsilon(-x_i)u^c_i\langle e,x\rangle\\ F_2(x,u) &= -\frac{\varrho}{2}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i) - \sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)x_i + \sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)u^c_i\langle e,x\rangle\end{aligned}\right\}\quad \forall\,(x,u)\in\mathcal{K}_0^+\times\Delta\,.$$

Again using (4.8), we have

$$\frac{1}{\theta}\,F_1(x,u) \le 1+\varrho\varepsilon-\varepsilon\|x^-\|_1 \qquad \forall\,(x,u)\in\mathcal{K}_0^+\times\Delta\,. \tag{4.24}$$

We break the estimate of F_2 in two parts. First, for any $\delta\in(0,1)$, using (4.8), we obtain

$$\begin{aligned}F_2(x,u) &\le -\frac{\varrho\varepsilon}{2}+1-\varepsilon\|x^+\|_1+\varepsilon\langle e,x\rangle\\ &\le -\frac{\varrho\varepsilon}{2}+1-\varepsilon\|x^-\|_1 \qquad \forall\,(x,u)\in\bigl(\mathcal{K}_0^+\setminus\mathcal{K}_\delta^+\bigr)\times\Delta\,.\end{aligned} \tag{4.25}$$

Combining (4.24) and (4.25), we get

$$\langle b(x,u), \nabla V(x)\rangle \le \biggl(\theta(1+\varrho\varepsilon) - \frac{\varepsilon(1-\delta)}{2}\|x\|_1\biggr)V(x) \qquad \forall\,(x,u)\in\bigl(\mathcal{K}_0^+\setminus\mathcal{K}_\delta^+\bigr)\times\Delta\,, \tag{4.26}$$

and $\theta \ge 1$, where we use the fact that $\|x^-\|_1 \ge \frac{1-\delta}{2}\|x\|_1$ on $\mathcal{K}_0^+\setminus\mathcal{K}_\delta^+$ by (4.10). Thus (4.14a) follows by (4.23) and (4.26).

Next, using (4.8), we have

$$F_2(x,u) \,\le\, -\frac{\varrho\varepsilon}{2} \qquad \forall\,(x,u)\in\mathcal{K}_\delta^+\times\Delta\,,$$

and this completes the proof. □

4.3.2. *The case of networks with a dominant server pool.* Consider the class of networks described in Subsection 2.4.1. We have the following lemma.

Lemma 4.3. *Consider a network with a dominant server pool, such that $\varrho>0$. Let $\delta\in(0,1)$, $\theta\ge\theta_0 := 2\,\frac{\max_i\mu_{i1}}{\min_i\mu_{i1}}$, and $V(x)$ be as in (4.7). Then, (4.14a) and (4.14b) hold.*

Proof. The method we follow is analogous to the proof of Lemma 4.2. Recall the definitions in (4.20). A straightforward calculation using (2.33) shows that

$$\begin{aligned}\frac{1}{\theta}\,F_1(x,u) \;&=\; \frac{\varrho}{m}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(-x_i) + \sum_{i\in\mathcal{I}}\psi'_\varepsilon(-x_i)\bigl(x_i - u_i^c\langle e,x\rangle^+\bigr) - \langle e,x\rangle^-\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}_1(i)}\psi'_\varepsilon(-x_i)(\eta_{ij}-1)u_j^s\,,\\ F_2(x,u) \;&=\; -\frac{\varrho}{m}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i) - \sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)\bigl(x_i - u_i^c\langle e,x\rangle^+\bigr) + \langle e,x\rangle^-\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}_1(i)}\psi'_\varepsilon(x_i)(\eta_{ij}-1)u_j^s\,.\end{aligned}$$

Let $\underline{\eta} := \min_{ij}\eta_{ij}$, and $\bar\eta := \max_{ij}\eta_{ij}$. Noting that

$$-\langle e,x\rangle^-\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}_1(i)}\psi'_\varepsilon(-x_i)(\eta_{ij}-1)u_j^s \,\le\, \varepsilon(1-\underline{\eta})^+\langle e,x\rangle^-\,,$$

it is easy to verify using (4.9) and (4.10) that

$$\begin{aligned}\frac{1}{\theta}\,F_1(x,u) \;&\le\; \varrho\varepsilon + \frac{m}{2} - \varepsilon\|x^-\|_1 + \varepsilon(1-\underline{\eta})^+\|x^-\|_1\\ &\le\; \varrho\varepsilon + \frac{m}{2} - \frac{\varepsilon(1-\delta)}{2}(\underline{\eta}\wedge 1)\|x\|_1 \qquad \forall (x,u)\in\bigl(\mathcal{K}_\delta^+\bigr)^{\mathsf{c}}\times\Delta\,.\end{aligned} \tag{4.27}$$

Note that the drift equations on $\mathcal{K}_0^+$ are similar to those of the 'N' model, with the only exception that $\frac{\varrho}{2}$ is replaced by $\frac{\varrho}{m}$, and the sum ranges from $i=1,\dots,m$ instead of $i=1,2$. Hence, we obtain

$$\begin{aligned}F_2(x,u) \;&\le\; -\frac{\varrho\varepsilon}{m} + \frac{m}{2} - \varepsilon\|x^+\|_1 + \varepsilon\langle e,x\rangle\\ &\le\; -\frac{\varrho\varepsilon}{m} + \frac{m}{2} - \varepsilon\|x^-\|_1\\ &\le\; -\frac{\varrho\varepsilon}{m} + \frac{m}{2} - \frac{\varepsilon(1-\delta)}{2}\|x\|_1 \quad \forall\,(x,u)\in\bigl(\mathcal{K}_0^+\setminus\mathcal{K}_\delta^+\bigr)\times\Delta\,,\end{aligned} \tag{4.28}$$

and

$$F_2(x,u) \;\le\; -\frac{\varrho\varepsilon}{m} - \varepsilon\langle e,x\rangle + \varepsilon\langle e,x\rangle \;\le\; -\frac{\varrho\varepsilon}{m} \qquad \forall (x,u)\in\mathcal{K}_\delta^+\times\Delta\,, \tag{4.29}$$

for any $\delta\in(0,1)$.

The choice of θ implies that $V_1\ge V_2^2$ on $\mathcal{K}_0^-$. Thus (4.14a) holds by (4.27) and (4.28), while (4.29) is equivalent to (4.14b). □

4.3.3. *The case of networks with class-dependent service rates.* Consider the class of networks described in Subsection 2.4.2. Such networks have a limiting diffusion with the same drift structure studied in [4], and that paper shows that when $\varrho>0$, then the diffusion (and the prelimit) is uniformly exponentially ergodic in the presence or absence of abandonment. However, the proof of uniform exponential ergodicity of the prelimit for models with class-dependent service rates does not seem to carry through with the Lyapunov function used in [4]. Thus for the sake of proving the result for the n^{th} system in Section 5, we adopt here the Lyapunov function in (4.7).

Lemma 4.4. *Consider a network satisfying $\mu_{ij} = \mu_i$ for all $i \in \mathcal{I}$, and $\varrho > 0$. Let $\delta \in (0,1)$, and $\theta \ge \theta_0 := 2\,\frac{\mu_{\max}}{\mu_{\min}}$. Then, (4.14a) and (4.14b) hold.*

Proof. A simple calculation using (2.34) shows that

$$\begin{aligned}\frac{1}{\theta}\,F_1(x,u) \;&=\; \frac{\varrho}{m}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(-x_i) + \sum_{i\in\mathcal{I}}\psi'_\varepsilon(-x_i)\bigl(x_i - u^c_i\langle e,x\rangle^+\bigr)\,,\\ F_2(x,u) \;&=\; -\frac{\varrho}{m}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i) - \sum_{i\in\mathcal{I}}\psi'_\varepsilon(x_i)\bigl(x_i - u^c_i\langle e,x\rangle^+\bigr)\,.\end{aligned} \tag{4.30}$$

Using (4.30), we obtain

$$\frac{1}{\theta}\,F_1(x,u) \;\le\; \varrho\varepsilon + \frac{m}{2} - \frac{\varepsilon}{2}\|x\|_1 \qquad \forall (x,u)\in\mathcal{K}_0^-\times\Delta\,,$$

Therefore, (4.14a) holds on $\mathcal{K}_0^-\times\Delta$ by this inequality and the choice of θ.

On $\mathcal{K}_0^+\times\Delta$, the equations in (4.30) are identical to the corresponding ones for a network with a dominant server pool, for which the result has already been established in Lemma 4.3. This completes the proof. □

5. Uniform exponential ergodicity of the n^{th} system

In this section we show that if $\varrho_n > 0$ then the prelimit of a network with a dominant server pool, or with class-dependent service rates, is uniformly exponentially ergodic and the invariant distributions have exponential tails.

Recall that $\{\tilde\mu_i\,,\ i\in\mathcal{I}\}$ are the elements of the diagonal matrix B^n_1 in (2.25). Throughout this section V denotes the function in (4.7), with ε given by

$$\varepsilon \;=\; \varepsilon_n \;:=\; \frac{\varrho_n}{3m}\left(\sum_{i\in\mathcal{I}}\frac{1}{n}\,\frac{\lambda^n_i(3\tilde\mu^n_i+2)}{(\tilde\mu^n_i)^2}\right)^{-1}\exp\left(-\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}}\frac{1}{\tilde\mu^n_i}\right). \tag{5.1}$$

Recall the definition of the operator $\widehat{\mathcal{L}}^n_{\hat z}$ in (2.15), and the definitions of $\mathcal{S}^n$ and $\widetilde{\mathcal{Z}}^n(\hat x)$ in (2.12) and Definition 2.1. We start with the following simple assertion.

Lemma 5.1. *Let V be the function in (4.7) with ε as in (5.1), and θ fixed at some value. Suppose that for any $\delta\in(0,1)$ there exist positive constants c_0 and c_1 such that the drift b^n in (2.25) satisfies*

$$\begin{aligned}\bigl\langle b^n(\hat x,\hat z),\nabla V(\hat x)\bigr\rangle \;&\le\; c_0 - \varepsilon c_1\|x\|_1 V(\hat x) &&\forall\,\hat x\in\mathcal{S}^n\setminus\mathcal{K}^+_\delta\,,\ \forall\,\hat z\in\widetilde{\mathcal{Z}}^n(\hat x)\,,\\ \bigl\langle b^n(\hat x,\hat z),\nabla V_2(\hat x)\bigr\rangle \;&\le\; -\frac{\varrho_n\varepsilon_n}{2m}V_2(\hat x) &&\forall\,\hat x\in\mathcal{S}^n\cap\mathcal{K}^+_\delta\,,\ \forall\,\hat z\in\widetilde{\mathcal{Z}}^n(\hat x)\,.\end{aligned} \tag{5.2}$$

Then, there exists a constant $\widehat C_0$ such that

$$\widehat{\mathcal{L}}^n_{\hat z}V(\hat x) \;\le\; \widehat C_0 - \frac{\varrho_n\varepsilon_n}{4m}V(\hat x) \qquad \forall\,\hat x\in\mathcal{S}^n\,,\ \forall\,\hat z\in\widetilde{\mathcal{Z}}^n(\hat x)\,. \tag{5.3}$$

Proof. A simple calculation shows that

$$\int_0^1(1-t)\,\partial_{x_ix_i}V_2\Bigl(\hat x\pm\tfrac{t}{\sqrt{n}}e_i\Bigr)\,\mathrm{d}t \;\le\; \frac{\varepsilon_n^2}{2}\left(\sum_{i\in\mathcal{I}}\frac{(3\tilde\mu^n_i+1)}{(2\tilde\mu^n_i)^2}\right)\exp\left(\frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}}\frac{1}{\tilde\mu^n_i}\right)V_2(\hat x)\,.$$

Thus, using (3.5) to express the first and second order incremental quotients in (2.15), we obtain

$$\widehat{\mathcal{L}}^n_{\hat z}V_2(\hat x) \;\le\; \frac{\varrho_n\varepsilon_n}{4m}V_2(\hat x) + \bigl\langle b^n(\hat x,\hat z),\nabla V_2(\hat x)\bigr\rangle\,.$$

The rest follows as in the proof of Lemma 4.1 by selecting δ sufficiently close to 1. □

Remark 5.1. Recall (2.25). In direct analogy to Remark 4.2, if we let $\zeta^n = \frac{\varrho_n}{m} e + (B_1^n)^{-1}\ell^n$, with ϱ_n as in (2.35), then a mere translation of the origin of the form $\tilde{X}^n = \hat{X}^n + \zeta^n$ results in a diffusion of with the same drift as (2.25), except that the vector ℓ^n gets replaced by $\ell^n = -\frac{\varrho_n}{m}(B_1)^n e$. Therefore, we may assume without any loss of generality that the drift in (2.25) takes the form

$$b^n(\hat{x}, \hat{z}) \;=\; -\frac{\varrho_n}{m} B_1^n e - B_1^n\bigl(\hat{x} - \langle e, \hat{x}\rangle^+ u^c\bigr) + B_2^n u^s \langle e, \hat{x}\rangle^- + \hat{\vartheta}^n(\hat{x}, \hat{z})\bigl(B_1^n u^c + B_2^n u^s\bigr)\,. \tag{5.4}$$

Note that this centering has the effect of translating the 'equilibrium' allocations $\overline{z}_{ij}^n$ given in (2.8). Since this translation is of $\mathcal{O}(n^{-1/2})$, it has no effect on the results for large n. However, in the interest of providing precise estimates we calculate the new values of $\overline{z}_{ij}^n$. Note that $\langle e, \zeta^n\rangle = 0$, and recall the map Φ in (2.22). Let $\breve{z}_{ij}^n = \Phi(\zeta^n, 0)$. Then, the centering of $\hat{x}$ that results in (5.4) is given by (compare with (2.8))

$$\bar{z}_{ij}^n \;=\; \frac{1}{n}\xi_{ij}^* N_j^n + \frac{\breve{z}_{ij}^n}{\sqrt{n}}\,, \qquad \bar{x}_i^n \;:=\; \sum_{j\in\mathcal{J}} \bar{z}_{ij}^n\,. \tag{5.5}$$

Throughout this section the family $\{\overline{z}_{ij}^n\ (i,j) \in \mathcal{E}\}$ is as given in (5.5).

Remark 5.2. Recall Definition 2.1 and (2.26). Let $\hat{z} \in \widetilde{\mathcal{Z}}^n(\hat{x})$. Then, $\hat{\vartheta}^n(\hat{x}, \hat{z}) = \hat{\vartheta}_*^n(\hat{x}) = 0$ for all $\hat{x} \in \breve{\mathfrak{X}}^n$, and in view of (5.4), for any $\hat{x} \in \breve{\mathfrak{X}}^n$, there exists $u = u(\hat{x}, \hat{z}) \in \mathit{\Delta}$ such that

$$b^n(\hat{x}, \hat{z}) \;=\; -\frac{\varrho_n}{m} B_1^n e - B_1^n\bigl(\hat{x} - \langle e, \hat{x}\rangle^+ u^c\bigr) + B_2^n u^s \langle e, \hat{x}\rangle^-\,. \tag{5.6}$$

In view of Lemma 5.1, and using Lemmas 4.2 to 4.4, it is clear that Foster–Lyapunov equations for $\mathcal{L}_u$ carry over to analogous equations for $\widehat{\mathcal{L}}_{\hat{z}}^n$ on $\breve{\mathfrak{X}}^n$ uniformly over SWC policies. However, even though $\breve{\mathfrak{X}}^n$ fills the whole space as $n \to \infty$, b and b^n differ in functional form when $\hat{\vartheta}^n(\hat{x}, \hat{z}) \neq 0$, and this makes the stability analysis of multiclass multi-pool networks much harder than the 'V' network studied in [4].

Notation 5.1. Let ε_n and $\bar{z}_{ij}^n$ as in (5.1) and (5.5), respectively. For a network with a dominant server pool as in Subsection 4.3.2 define

$$n_0 \;:=\; \max\left\{n \in \mathbb{N} : \frac{1}{\sqrt{n}} \geq \varepsilon_n \min_{i\in\mathcal{I}} \bar{z}_{i1}^n\right\}, \tag{5.7}$$

while for network with class-dependent service rates, we let

$$n_0 \;:=\; \max\left\{n \in \mathbb{N} : \frac{1}{\sqrt{n}} \geq \frac{\varepsilon_n}{2m} \min_{i\sim j} \bar{z}_{ij}^n\right\}. \tag{5.8}$$

Since $\{\varepsilon_n\}$ and $\{\bar{z}_{i1}^n\}$ are bounded away from 0 by the convergence of the parameters in (2.2), the number n_0 is finite.

The next theorem is the main result for the uniform exponential ergodicity of the prelimit processes. Recall (2.35). Notice the similarities between the results in Theorem 5.1 for the diffusion-scaled processes and the results in Theorem 4.2 and Remark 4.1 for the diffusion limit of the networks under consideration.

Theorem 5.1. *Assume that $\varrho_n > 0$, and let n_0 be as in Notation 5.1. Then the prelimit dynamics of any network with a dominant server pool or with class-dependent service rates are uniformly exponentially ergodic and the invariant distributions have exponential tails for all $n > n_0$. In particular, due to the convergence of the parameters, there exists $\widehat{C}_0$ independent of n such that*

$$\widehat{\mathcal{L}}_{\hat{z}}^n V(\hat{x}) \;\leq\; \widehat{C}_0 - \frac{\varrho_n \varepsilon_n}{4m} V(\hat{x}) \qquad \forall\, \hat{x} \in \mathcal{S}^n\,,\ \forall\, \hat{z} \in \widetilde{\mathcal{Z}}^n(\hat{x})\,. \tag{5.9}$$

where V and ε are as in (4.7) and (5.1).

In addition, with $P_t^{n,\hat{z}}$ and $\uppi_{\hat{z}}^n$ denoting respectively the transition probability and the stationary distribution of $\hat{X}^n(t)$ under a policy $\hat{z} \in \widetilde{\mathfrak{Z}}^n$, there exist positive constants γ and C_γ not depending on $n \geq 0$ or $\hat{z}$, such that

$$\big\|P_t^{n,\hat{z}}(\hat{x},\cdot\,) - \uppi_{\hat{z}}^n(\cdot)\,\big\|_V \;\leq\; C_\gamma V(\hat{x})\,\mathrm{e}^{-\gamma t}\,, \qquad \forall\,\hat{x}\in\mathcal{X}^n\,,\ \forall\,t\geq 0\,. \tag{5.10}$$

Proof. In Lemmas 5.4 and 5.5 in the section which follows, we establish (5.2) for these networks. Thus the proof of (5.9) follows directly from Lemma 5.1.

Since the process $\hat{X}^n$ is irreducible and aperiodic under any stationary Markov scheduling $\hat{z} \in \widetilde{\mathfrak{Z}}^n$ (see Definition 2.1), a convergence property completely analogous to (4.3) follows from (5.9). The proof of this fact is identical to [4, Theorem 2.1(b)]. □

Remark 5.3. Using (4.18) and (5.10), it is clear that under any scheduling policy $\hat{z} \in \widetilde{\mathfrak{Z}}^n$ with a corresponding control v, the stationary distribution of the diffusion-scaled process $\hat{X}^n(t)$ converges to that of the limiting diffusion $\{X_t\}_{t\geq 0}$ for the two classes of networks, that is,

$$\uppi_{\hat{z}}^n(\cdot) \;\to\; \uppi_v(\cdot)\,, \quad \text{as} \quad n\to\infty. \tag{5.11}$$

That is, the interchange of limits property holds.

5.1. **Four technical lemmas.** In this section, we establish the technical results used in the proof of Theorem 5.1.

Let

$$\widetilde{\mathcal{X}}^n \;:=\; \big\{\hat{x}\in\mathcal{S}^n : \hat{\vartheta}_*^n(\hat{x}) \neq 0\big\}\,, \tag{5.12}$$

with $\hat{\vartheta}_*^n$ as in Definition 2.1. As seen in Subsection 2.3, the set $\breve{\mathcal{X}}^n$ in (2.26) is contained in $\mathcal{S}^n \setminus \widetilde{\mathcal{X}}^n$. In establishing (5.2) on $\mathcal{S}^n \setminus \widetilde{\mathcal{X}}^n$, the results in Section 4 pave the way, since the drift of of the controlled generator $\widehat{\mathcal{L}}_{\hat{z}}^n$ over the class of SWC stationary Markov policies $\widetilde{\mathfrak{Z}}^n$ (see (5.6)) has the same functional form as the drift of the diffusion in (2.31). So it remains to establish (5.2) in $\widetilde{\mathcal{X}}^n$. We start by establishing a bound for $\hat{\vartheta}^n$ in (2.18) over all SWC policies.

As done earlier in the interest of notational economy, we suppress the dependence on n in the diffusion scaled variables $\hat{x}^n$ and $\hat{z}^n$ in (2.11).

Lemma 5.2. *There exists a number $\varkappa_\circ^n < 1$ depending only on the parameters of the network such that*

$$\hat{\vartheta}^n(\hat{x},\hat{z}) \;=\; \hat{\vartheta}_*^n(\hat{x}) \;\leq\; \varkappa_\circ^n\big(\|\hat{x}^+\|_1 \wedge \|\hat{x}^-\|_1\big) \quad \forall\,\hat{z}\in\widetilde{\mathcal{Z}}^n(\hat{x})\,. \tag{5.13}$$

In addition, due to the convergence of the parameters in (2.2), such a constant $\varkappa_\circ < 1$ may be selected which does not depend on n.

Before proceeding to the proof of Lemma 5.2, we provide an interpretation of (5.13). Recall from (2.10) that $\hat{x}_i = \frac{1}{\sqrt{n}}(x_i - \sum_{j\in\mathcal{J}} \xi_{ij}^* N_j^n)$, and observe that $\hat{x}_i$ is positive if the total number of class i customers exceeds the total number of servers assigned to class i in the fluid equilibrium $\big(\sum_{j\in\mathcal{J}} \xi_{ij}^* N_j^n\big)$ and is negative otherwise. This means that the quantity $\|\hat{x}^+\|_1 \wedge \|\hat{x}^-\|_1$ on the right hand side of (5.13) represents the minimum of the total number of customers in the queues and the total number of idle servers if the servers are assigned to customer classes according to the fluid equilibrium. Recall also that $\hat{\vartheta}_*^n$ represents the minimum of the total number of customers in the queues and the total number of idle servers under a SWC policy. The result in Lemma 5.2 is now clear, that is, the minimum of the total number of customers in the queues and idle servers is smaller under a SWC policy compared to a policy that assigns servers according to the fluid equilibrium.

Proof. Let $\hat{x} \in \widetilde{\mathfrak{X}}^n$, $\hat{z} \in \widetilde{\mathcal{Z}}^n(\hat{x})$, and define

$$\widetilde{\mathcal{J}} := \Biggl\{ j \in \mathcal{J} : \sum_{j \in \mathcal{J}(i)} \hat{z}_{ij} < 0 \Biggr\}, \quad \text{and } \widetilde{\mathcal{I}} := \bigl\{ i \in \mathcal{I} : (i,j) \in \mathcal{E} \text{ for some } j \in \widetilde{\mathcal{J}} \bigr\},$$

and $\widetilde{\mathcal{E}} := \bigl\{(i,j) \in \mathcal{E} : (i,j) \in \widetilde{\mathcal{I}} \times \widetilde{\mathcal{J}}\bigr\}$. Work conservation implies that $x_i^n = \sum_{j \in \mathcal{J}(i)} z_{ij}^n$ for all $i \in \widetilde{\mathcal{I}}$. Let $\hat{\imath} \in \mathcal{I}$ be such that $\hat{q}_{\hat{\imath}} > 0$, and consider the unique path (since the graph of the network is a tree) connecting $\hat{\imath}$ and $\widetilde{\mathcal{I}}$, that is, a path $\hat{\imath} \to j_1 \to i_1 \to j_2 \ldots \to j_k \to \tilde{\imath}$, with $j_\ell \in \mathcal{J} \setminus \widetilde{\mathcal{J}}$ for $\ell = 1, \ldots, k$, $i_\ell \in \mathcal{I} \setminus \widetilde{\mathcal{I}}$ for $\ell = 1, \ldots, k-1$, and $\tilde{\imath} \in \widetilde{\mathcal{I}}$. We claim that $z^n_{\tilde{\imath}, j_k} = 0$, or equivalently, that $\hat{z}_{\tilde{\imath}, j_k} = -\bar{z}^n_{ij}/\sqrt{n}$, with $\bar{z}^n_{ij}$ as defined in (5.5). If not, then we can move a job of class $\tilde{\imath}$ from pool j_k to some pool in $\widetilde{\mathcal{J}}$, and proceeding along the path to place one additional job from class $\hat{\imath}$ into service, thus contradicting the hypothesis that $\hat{z} \in \widetilde{\mathcal{Z}}^n(\hat{x})$. Removing all such paths, we are left with a strict subnetwork (possibly disconnected) $\mathcal{G}_\circ = \bigl(\mathcal{I}_\circ \cup \mathcal{J}_\circ, \mathcal{E}_\circ\bigr)$, with $\mathcal{I}_\circ \supset \widetilde{\mathcal{I}}$, $\mathcal{J}_\circ \supset \widetilde{\mathcal{J}}$, and $\mathcal{E}_\circ := \bigl\{(i,j) \in \mathcal{E} : (i,j) \in \mathcal{I}_\circ \times \mathcal{J}_\circ\bigr\}$, such that

$$x_i^n = \sum_{j \in \mathcal{J}(i) \cap \mathcal{J}_\circ} z_{ij}^n, \quad \forall\, i \in \mathcal{I}_\circ. \tag{5.14}$$

Let $\mathcal{E}'_\circ := \bigl(\mathcal{I}_\circ \times (\mathcal{J} \setminus \mathcal{J}_\circ)\bigr) \cap \mathcal{E}$. By (5.14) we have

$$\sum_{(i,j) \in \mathcal{E}'_\circ} z_{ij}^n \,=\, 0\,.$$

Thus we have

$$\|\hat{x}^-\|_1 \,\ge\, -\sum_{i \in \mathcal{I}_\circ} \hat{x}_i \,=\, \sqrt{n} \sum_{(i,j) \in \mathcal{E}'_\circ} \bar{z}_{ij}^n - \sum_{(i,j) \in \mathcal{E}_\circ} \hat{z}_{ij} \,\ge\, \sqrt{n} \sum_{(i,j) \in \mathcal{E}'_\circ} \bar{z}_{ij}^n + \hat{\vartheta}_*^n(\hat{x}) \tag{5.15}$$

by the construction above. By (5.15), we obtain

$$\begin{aligned} \hat{\vartheta}_*^n(\hat{x}) \,&\le\, \frac{\sum_{(i,j) \in \mathcal{E}_\circ} \hat{z}_{ij}}{\sqrt{n} \sum_{(i,j) \in \mathcal{E}'_\circ} \bar{z}_{ij}^n - \sum_{(i,j) \in \mathcal{E}_\circ} \hat{z}_{ij}} \sum_{i \in \mathcal{I}_\circ} \hat{x}_i \\ &\le\, -\frac{\sum_{(i,j) \in \mathcal{E}_\circ} \bar{z}_{ij}^n}{\sum_{(i,j) \in \mathcal{E}_\circ} \bar{z}_{ij}^n + \sum_{(i,j) \in \mathcal{E}'_\circ} \bar{z}_{ij}^n} \sum_{i \in \mathcal{I}_\circ} \hat{x}_i\,. \end{aligned} \tag{5.16}$$

Similarly,

$$\|\hat{x}^+\|_1 \,\ge\, \sum_{i \in \mathcal{I} \setminus \mathcal{I}_\circ} \hat{x}_i \,\ge\, \sqrt{n} \sum_{(i,j) \in \mathcal{E}'_\circ} \bar{z}_{ij}^n + \hat{\vartheta}_*^n(\hat{x})\,. \tag{5.17}$$

Using the bound $\hat{\vartheta}_*^n(\hat{x}) \le \sqrt{n} \sum_{(i,j) \in \mathcal{E}_\circ} \bar{z}_{ij}^n$ we obtain from (5.17) that

$$\begin{aligned} \hat{\vartheta}_*^n(\hat{x}) \,&\le\, \frac{\Bigl(\sum_{i \in \mathcal{I} \setminus \mathcal{I}_\circ} \hat{x}_i - \sqrt{n} \sum_{(i,j) \in \mathcal{E}'_\circ} \bar{z}_{ij}^n\Bigr) \wedge \sqrt{n} \sum_{(i,j) \in \mathcal{E}_\circ} \bar{z}_{ij}^n}{\sum_{i \in \mathcal{I} \setminus \mathcal{I}_\circ} \hat{x}_i} \sum_{i \in \mathcal{I} \setminus \mathcal{I}_\circ} \hat{x}_i \\ &\le\, \frac{\sum_{(i,j) \in \mathcal{E}_\circ} \bar{z}_{ij}^n}{\sum_{(i,j) \in \mathcal{E}_\circ} \bar{z}_{ij}^n + \sqrt{n} \sum_{(i,j) \in \mathcal{E}'_\circ} \bar{z}_{ij}^n} \sum_{i \in \mathcal{I} \setminus \mathcal{I}_\circ} \hat{x}_i\,. \end{aligned} \tag{5.18}$$

It should be now clear how to construct $\varkappa_\circ^n$. For any given subset $\mathcal{J}' \subsetneq \mathcal{J}$, let

$$\mathcal{I}_{\mathcal{J}'} \,:=\, \cup_{j \in \mathcal{J}'} \mathcal{I}(j)\,, \qquad \mathcal{E}_{\mathcal{J}'} \,:=\, \bigl\{(i,j) \in \mathcal{E} : (i,j) \in \mathcal{I}_{\mathcal{J}'} \times \mathcal{J}'\bigr\}\,,$$

and $\mathcal{E}'_{\mathcal{J}'} := \bigl(\mathcal{I}_{\mathcal{J}'} \times (\mathcal{J} \setminus \mathcal{J}')\bigr) \cap \mathcal{E}$, and define

$$\varkappa^n_\circ := \max_{\mathcal{J}' \subsetneq \mathcal{J}} \frac{\sum_{(i,j)\in\mathcal{E}_{\mathcal{J}'}} \bar{z}^n_{ij}}{\sum_{(i,j)\in\mathcal{E}_{\mathcal{J}'}} \bar{z}^n_{ij} + \sum_{(i,j)\in\mathcal{E}'_{\mathcal{J}'}} \bar{z}^n_{ij}}\,.$$

Then the result clearly follows from (5.16) and (5.18) since

$$\|\hat{x}^-\|_1 \,\ge\, -\sum_{i\in\mathcal{I}_\circ} \hat{x}_i\,, \quad \text{and} \quad \|\hat{x}^+\|_1 \,\ge\, \sum_{i\in\mathcal{I}\setminus\mathcal{I}_\circ} \hat{x}_i\,.$$

This completes the proof. □

Remark 5.4. We provide an example of a SWC policy for the 'N' network with two classes of customers and two server pools where class 1 can be served by both server pools while class 2 can only served by pool 2 to explain the analysis in the proof of Lemma 5.2. The SWC policy is given by

$$\begin{aligned}
z_{11}(x) \;&=\; x_1 \wedge N^n_1 \\
z_{12}(x) \;&=\; \begin{cases} (x_1 - N^n_1)^+ \wedge \xi^*_{12} N^n_2 & \text{if } x_2 \ge \xi^*_{22} N^n_2 \\ (x_1 - N^n_1)^+ \wedge (N^n_2 - x_2) & \text{otherwise,} \end{cases} \\
z_{22}(x) \;&=\; \begin{cases} x_2 \wedge \xi^*_{22} N^n_2 & \text{if } x_1 \ge N^n_1 + \xi^*_{12} N^n_2 \\ x_2 \wedge \bigl(N^n_2 - (x_1 - N^n_1)^+\bigr) & \text{otherwise.} \end{cases}
\end{aligned}$$

This is a priority policy in which customer class 1 prefers pool 1 over pool 2. This means that customer class 1 can use servers in pool 2 only if there are no idle servers in pool 1.

Recall from (2.9) and (2.10) that $\hat{q}_i \ge 0$ for all $i \in \mathcal{I}$. Recall also that we are only considering work-conserving policies and that $\hat{\vartheta}^n$ represents the minimum of the total number of customers in the queues and the total number of idle servers. For the 'N' network with two classes of customers and two server pools, we have the following cases:

Case 1 : $\hat{q}_1 > 0$ and $\hat{q}_2 \ge 0$, in which case we have no idle servers and hence $\hat{\vartheta}^n_*(\hat{x}) = 0$.

Case 2 : $\hat{q}_1 = 0$ and $\hat{q}_2 = 0$. Again, this is a trivial case and means that we have no customers in the queues which implies that $\hat{\vartheta}^n_*(\hat{x}) = 0$.

Case 3 : $\hat{q}_1 = 0$ and $\hat{q}_2 > 0$. This is actually the case that requires some analysis. Here again we have the following cases:

- a : $\hat{y}_1 = 0$ which means that there are no idle servers in pool 1 and hence $\hat{\vartheta}^n_*(\hat{x}) = 0$. This is because $\hat{q}_2 > 0$ which implies that $\hat{y}_2 = 0$ under work conservation.
- b : $\hat{y}_1 > 0$ which means that there are some idle servers in pool 1. This is the case analyzed in the proof of Lemma 5.2. (Observe the following notation in the proof: $\widetilde{\mathcal{J}} = \{1\}$, $\widetilde{\mathcal{I}} = \{1\}$, $\hat{\imath} = 2$, the path is class 2 → pool 2 → class 1, and $\mathcal{G}_\circ = \{(1,1)\}$.) Note that since we are using a SWC policy, this means that no class 1 customers are being served in pool 2 because otherwise, one can move a class 1 customer from pool 2 to pool 1 and get a smaller $\hat{\vartheta}^n$. Hence,

$$\hat{\vartheta}^n_*(\hat{x}) = \frac{1}{\sqrt{n}}\bigl((N^n_1 - x_1) \wedge (x_2 - N^n_2)\bigr)\,,$$

  where $(N^n_1 - x_1)$ is the total number of idle server and $(x_2 - N^n_2)$ is the total number of customers in the queues. This is because the only pool having idle servers is pool 1 and the only class having customers in the queue is class 2 ($\hat{q}_1 = 0$).

Recall that $\hat{x}_i = \frac{1}{\sqrt{n}}(x_i - \sum_{j\in\mathcal{J}} \xi^*_{ij} N^n_j)$. This means in this case that

$$\hat{x}_1^+ = 0 \qquad ; \qquad \hat{x}_1^- = \frac{1}{\sqrt{n}}(N_1^n + \xi^*_{12}N_2^n - x_1)$$

$$\hat{x}_2^+ = \frac{1}{\sqrt{n}}(x_2 - \xi^*_{22}N_2^n) \qquad ; \qquad \hat{x}_2^- = 0\,.$$

Therefore, we have the following equation

$$\|\hat{x}^+\|_1 \wedge \|\hat{x}^-\|_1 = \frac{1}{\sqrt{n}}\Big((N_1^n + \xi^*_{12}N_2^n - x_1) \wedge (x_2 - \xi^*_{22}N_2^n)\Big).$$

Note also that (2.4) ($\xi^*_{12} + \xi^*_{22} = 1$) implies that

$$N_1^n - x_1 \le x_2 - N_2^n \iff N_1^n + \xi^*_{12}N_2^n - x_1 \le x_2 - \xi^*_{22}N_2^n\,.$$

It is now clear that there exists a constant $\varkappa^n_\circ < 1$ such that $\hat{\vartheta}^n_* < \varkappa^n_\circ\big(\|\hat{x}^+\|_1 \wedge \|\hat{x}^-\|_1\big)$ where

$$\varkappa^n_\circ = \frac{N_1^n - x_1}{N_1^n + \xi^*_{12}N_2^n - x_1} \vee \frac{x_2 - N_2^n}{x_2 - \xi^*_{22}N_2^n}.$$

This completes the analysis of all the cases.

Remark 5.5. It is easy to see that the estimates of the bounds on $\hat{\vartheta}^n$ can be improved. It is clear from (5.15) and (5.16), that

$$\hat{\vartheta}^n_*(\hat{x}) \le \left(-\varkappa^n_0 \sum_{i\in\mathcal{I}_0} \hat{x}_i\right) \wedge \left(\sum_{i\in\mathcal{I}\setminus\mathcal{I}_\circ} \hat{x}_i\right) \qquad \forall\, \hat{x} \in \widetilde{\mathfrak{X}}^n\,.$$

Also, since there can be at most $\sum_{j\in\mathcal{J}} N^n_j$ idle servers, it follows that $\widetilde{\varkappa}^n_\circ \in (0,1)$, such that

$$-\sum_{i\in\mathcal{I}_0} \hat{x}_i \ge \widetilde{\varkappa}^n_\circ \|\hat{x}^-\|_1 \qquad \forall\, \hat{x} \in \widetilde{\mathfrak{X}}^n\,,$$

where the constant $\widetilde{\varkappa}^n_\circ \in (0,1)$ can be selected as

$$\widetilde{\varkappa}^n_\circ := \left(\sum_{j\in\mathcal{J}} N^n_j\right)^{-1} \min_{(i,j)\,\in\,\mathcal{E}} \xi^*_{ij} N^n_j\,.$$

Due to the convergence of the parameters in (2.2), $\widetilde{\varkappa}^n_\circ$ is bounded away from 0 uniformly in $n \in \mathbb{N}$. Note that for the 'N' network this translates to

$$\widetilde{\varkappa}^n_\circ = \frac{N_1^n \wedge \xi^*_{12}N_2^n \wedge \xi^*_{22}N_2^n}{N_1^n + N_2^n}\,.$$

Even though the 'N' network is a special case of networks with a dominant server pool we first establish the result for this network in Lemma 5.3 in order to exhibit with simpler calculations how Lemma 5.2 is applied.

Throughout the proofs of Lemmas 5.3 to 5.5 we use the functions (compare with (4.20))

$$F^n_i(\hat{x}, \hat{z}) := \frac{1}{V_i(x)} \big\langle b^n(\hat{x}, \hat{z}), \nabla V_1(\hat{x})\big\rangle\,, \qquad i = 1, 2\,,$$

and let n_0 be as in Notation 5.1. Moreover, we suppress the dependence on n in the variables $\hat{q}^n$, $\hat{y}^n$, and $\hat{\vartheta}^n$ in (2.17) and (2.18), and from ε_n in (5.1).

5.1.1. *The diffusion-scale of the 'N' network.* We recall here [27]. Recall also that we label the non-leaf server node as $j = 1$ without loss of generality and hence we present Stolyar's work in [27] using our notation. In this work, Stolyar considers the 'N' network with $\mathcal{O}(\sqrt{n})$ safety staffing in pool 1, under the priority discipline that class 2 has priority in pool 1 and class 1 prefers pool 2, and shows tightness of the invariant distributions. First note that for any stationary Markov scheduling policy z, such that class 2 has priority in pool 1 we have $z^n_{21}(x) = x^n_2 \wedge N^n_1$, and it is clear that such a policy is SWC. The same applies to Markov policies under which class 1 prefers pool 2 (here $z^n_{12}(x) = x^n_1 \wedge N^n_2$). As a result, SWC policies are more general than the particular policy considered in [27]. Recall that the matrices B^n_1 and B^n_2 in the drift (5.4) are given by

$$B^n_1 \;=\; \begin{pmatrix} \mu^n_{11} & 0 \\ 0 & \mu^n_{21} \end{pmatrix}\,, \qquad \text{and} \;\; B^n_2 \;=\; \begin{pmatrix} 0 & \mu^n_{12} - \mu^n_{11} \\ 0 & 0 \end{pmatrix} \tag{5.19}$$

It is also worth noting here, that the spare capacity ϱ_n of the n^{th} system is given by

$$\varrho_n \;=\; -\frac{1}{\sqrt{n}}\biggl(\frac{\lambda^n_1}{\mu^n_{11}} + \frac{\lambda^n_2}{\mu^n_{21}} - \frac{\mu^n_{12}N^n_2 + \mu^n_{11}\xi^*_{11}N^n_1}{\mu^n_{11}} - \frac{\mu^n_{21}\xi^*_{21}N^n_1}{\mu^n_{21}}\biggr)\,,$$

with

$$\xi^*_{11} \;=\; \frac{\lambda_1 - \mu_{12}\nu_2}{\mu_{11}\nu_1}\,, \quad \text{and} \quad \xi^*_{21} \;=\; \frac{\lambda_2}{\mu_{21}\nu_1}\,.$$

This is clear by (2.14), (2.35), and (5.19), together with [6, Equation (2)]. We let $\eta^n := \frac{\mu^n_{12}}{\mu^n_{11}}$.

Lemma 5.3. *Consider the 'N' network, and assume that $\varrho_n > 0$. Then for any $\theta \ge \theta^n_0 := \frac{\mu^n_{11} \vee \mu^n_{21}}{\mu^n_{11} \wedge \mu^n_{21}}$, and $\delta \in (0,1)$, there exist positive constants c_0 and c_1 such that (5.2) holds for all $n \ge n_0$.*

Proof. Recall here that $j = 1$ is the non-leaf server pool. As discussed earlier, it suffices to establish (5.2) in $\widetilde{\mathcal{X}}^n$. It is clear that $\hat{x}^-_1 = \hat{y}_2 + \sqrt{n}\hat{z}_{11}$, and $\hat{x}_2 = \hat{q}_2 + \sqrt{n}\hat{z}_{11}$ for all $\hat{z} \in \hat{\mathcal{Z}}^n(\hat{x})$ and $\hat{x} \in \widetilde{\mathcal{X}}^n$, with $\widetilde{\mathcal{X}}^n$ as defined in (5.12). Hence $u^c_1 = 0$, $u^c_2 = 1$, $u^s_2 = 1$, and $u^s_1 = 0$. Note here that SWC policies are interpreted through u^c and u^s as follows: idle servers are only allowed in pool 1 and customer queues are only allowed in class 2 which means that class 1 must use all the servers in pool 2 before using servers in pool 1. Also, by the definitions of ψ_ε and n_0 we have

$$\psi'_\varepsilon(\hat{x}_1) \;=\; \psi'_\varepsilon(-\hat{x}_2) \;=\; 0 \qquad \forall\, \hat{x} \in \widetilde{\mathcal{X}}^n\,,\ \forall\, n \ge n_0\,. \tag{5.20}$$

By (5.4) and (5.19), we have

$$\begin{aligned}\frac{1}{\theta}\,F^n_1(\hat{x},\hat{z}) \;=\; \frac{\varrho_n}{2}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(-\hat{x}_i) + \sum_{i\in\mathcal{I}}\psi'_\varepsilon(-\hat{x}_i)\bigl(\hat{x}_i - u^c_i\langle e,\hat{x}\rangle^+\bigr) - \psi'_\varepsilon(-\hat{x}_1)(\eta^n - 1)\langle e,\hat{x}\rangle^- \\ - \hat{\vartheta}^n\Bigl(\psi'_\varepsilon(-\hat{x}_2) + \psi'_\varepsilon(-\hat{x}_1)(\eta^n - 1)\Bigr)\,,\end{aligned} \tag{5.21}$$

$$\begin{aligned}F^n_2(\hat{x},\hat{z}) \;=\; -\frac{\varrho_n}{2}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(\hat{x}_i) - \sum_{i\in\mathcal{I}}\psi'_\varepsilon(\hat{x}_i)\bigl(\hat{x}_i - u^c_i\langle e,\hat{x}\rangle^+\bigr) + \psi'_\varepsilon(\hat{x}_1)(\eta^n - 1)\langle e,\hat{x}\rangle^- \\ + \hat{\vartheta}^n\Bigl(\psi'_\varepsilon(\hat{x}_2) + \psi'_\varepsilon(\hat{x}_1)(\eta^n - 1)\Bigr)\,.\end{aligned} \tag{5.22}$$

Using the fact that $\hat{x}_1^- - \hat{x}_2 = \hat{y}_2 - \hat{q}_2$, and $\hat{\vartheta}^n = \hat{q}_2$ when $\langle e, \hat{x}\rangle \le 0$, we obtain from (5.20) and (5.21) that

$$\begin{aligned}
\frac{1}{\theta}\, F_1^n(\hat{x},\hat{z}) \;&=\; \frac{\varrho_n}{2}\psi_\varepsilon'(-\hat{x}_1) - \psi_\varepsilon'(-\hat{x}_1)\hat{x}_1^- - \psi_\varepsilon'(-\hat{x}_1)(\eta^n-1)\langle e,\hat{x}\rangle^- - \psi_\varepsilon'(-\hat{x}_1)(\eta^n-1)\hat{\vartheta}^n\\
&= \varepsilon\Big(\frac{\varrho_n}{2} - \hat{x}_1^- - (\eta^n-1)\big(\hat{x}_1^- - \hat{x}_2\big) - (\eta^n-1)\hat{\vartheta}^n\Big)\\
&= \varepsilon\Big(\frac{\varrho_n}{2} - \hat{x}_1^- - (\eta^n-1)\hat{y}_2\Big)\\
&\le \varepsilon\Big(\frac{\varrho_n}{2} - \hat{x}_1^- + (1-\eta^n)^+\hat{x}_1\Big)\\
&\le \varepsilon\Big(\frac{\varrho_n}{2} - (\eta^n\wedge 1)\hat{x}_1^-\Big)\\
&\le \varepsilon\frac{\varrho_n}{2} - \frac{\varepsilon}{2}(\eta^n\wedge 1)\|\hat{x}\|_1 \qquad \forall\,(\hat{x},\hat{z}) \in \big(\widetilde{\mathfrak{X}}^n\cap\mathcal{K}_0^-\big)\times\hat{\mathcal{Z}}^n(\hat{x})\,,\ \forall\, n\ge n_0\,.
\end{aligned}\tag{5.23}$$

Similarly, from (5.22), we obtain

$$\begin{aligned}
F_2^n(\hat{x},\hat{z}) \;&=\; -\frac{\varrho_n}{2}\sum_{i\in\mathcal{I}}\psi_\varepsilon'(\hat{x}_i) - \sum_{i\in\mathcal{I}}\psi_\varepsilon'(\hat{x}_i)\hat{x}_i + \psi_\varepsilon'(\hat{x}_1)(\eta^n-1)\langle e,\hat{x}\rangle^-\\
&\qquad\qquad + \hat{\vartheta}^n\Big(\psi_\varepsilon'(\hat{x}_2) + \psi_\varepsilon'(\hat{x}_1)(\eta^n-1)\Big)\\
&\le\; 1 - \varepsilon\|\hat{x}^+\|_1 + \varepsilon\varkappa_0\Big(\|\hat{x}^-\|_1\wedge\|\hat{x}^+\|_1\Big)\\
&\le\; 1 - \varepsilon(1-\varkappa_0)\|\hat{x}^+\|_1 \qquad \forall\,(\hat{x},\hat{z}) \in \big(\widetilde{\mathfrak{X}}^n\cap\mathcal{K}_0^-\big)\times\hat{\mathcal{Z}}^n(\hat{x})\,,\ \forall\, n\ge n_0\,,
\end{aligned}\tag{5.24}$$

where we also use (4.8) and Lemma 5.2.

We continue with the estimate on $\mathcal{K}_0^+$. We have

$$\begin{aligned}
\frac{1}{\theta}\, F_1^n(\hat{x},\hat{z}) \;&\le\; \frac{\varrho_n}{2}\psi_\varepsilon'(-\hat{x}_1) - \psi_\varepsilon'(-\hat{x}_1)\hat{x}_1^- - \psi_\varepsilon'(-\hat{x}_1)(\eta^n-1)\hat{\vartheta}^n\\
&= \varepsilon\Big(\frac{\varrho_n}{2} - \hat{x}_1^- - (\eta^n-1)\hat{\vartheta}^n\Big)\\
&\le \varepsilon\Big(\frac{\varrho_n}{2} - (\eta^n\wedge 1)\hat{x}_1^-\Big) \qquad \forall\,(\hat{x},\hat{z}) \in \big(\widetilde{\mathfrak{X}}^n\cap\mathcal{K}_0^+\big)\times\hat{\mathcal{Z}}^n(\hat{x})\,,\ \forall\, n\ge n_0\,,
\end{aligned}\tag{5.25}$$

where in the last inequality we also use Lemma 5.2.

We break the estimate of F_2^n in two parts. First, using (4.8), (4.10), (5.20) and Lemma 5.2, we obtain

$$\begin{aligned}
F_2^n(\hat{x},\hat{z}) \;&\le\; -\frac{\varrho_n\varepsilon}{2} - \varepsilon\hat{x}_2 + \varepsilon\langle e,\hat{x}\rangle + \varepsilon\varkappa_0\Big(\hat{x}_1^-\wedge\hat{x}_2\Big)\\
&\le -\frac{\varrho_n\varepsilon}{2} - \varepsilon(1-\varkappa_0)\hat{x}_1^-\\
&\le \begin{cases} -\frac{\varrho_n\varepsilon}{2} - \frac{\varepsilon(1-\delta)}{2}(1-\varkappa_0)\|\hat{x}\|_1 & \text{for } (\hat{x},\hat{z})\in\big(\widetilde{\mathfrak{X}}^n\cap(\mathcal{K}_0^+\setminus\mathcal{K}_\delta^+)\big)\times\hat{\mathcal{Z}}^n(\hat{x})\\ -\frac{\varrho_n\varepsilon}{2} & \text{for } (\hat{x},\hat{z})\in\big(\widetilde{\mathfrak{X}}^n\cap\mathcal{K}_\delta^+\big)\times\hat{\mathcal{Z}}^n(\hat{x}). \end{cases}
\end{aligned}\tag{5.26}$$

Thus, (5.2) follows by (4.10) and (5.23)–(5.26). This completes the proof. □

5.1.2. *The diffusion scale of networks with a dominant pool.* We describe these networks exactly as in Subsection 4.3.2 where the dominant server pool is $j = 1$. We first note that the spare capacity ϱ_n of the n^{th} system is given by

$$\varrho_n \;=\; -\frac{1}{\sqrt{n}}\left(\sum_{i\in\mathcal{I}}\frac{\lambda_i^n}{\mu_{i1}^n} - \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}(i)}\frac{\mu_{ij}^n}{\mu_{i1}^n}\xi_{ij}^* N_j^n\right),$$

where ξ^*_{ij} satisfies

$$\sum_{j\in\mathcal{J}(i)} \mu_{ij}\xi^*_{ij}\nu_j = \lambda_i\,.$$

This is again due to (2.4), (2.14), (2.34), and (2.35).

Recall from (5.4) that the drift reduces to the following form:

$$\begin{aligned} b^n_i(\hat{x},\hat{z}) \;=\; &-\frac{\varrho_n}{m}\mu^n_{i1} - \mu^n_{i1}\bigl(\hat{x}_i - u^c_i\langle e,\hat{x}\rangle^+\bigr) + \sum_{j\in\mathcal{J}_1(i)} \mu^n_{i1}\bigl(\eta^n_{ij}-1\bigr)u^s_j\langle e,\hat{x}\rangle^- \\ &+ \hat{\vartheta}^n\Biggl(\mu^n_{i1}u^c_i + \sum_{j\in\mathcal{J}_1(i)} \mu^n_{i1}\bigl(\eta^n_{ij}-1\bigr)u^s_j\Biggr)\,, \qquad i\in\mathcal{I}\,, \end{aligned} \tag{5.27}$$

with $\eta^n_{ij} \coloneqq \frac{\mu^n_{ij}}{\mu^n_{i1}}$ for $j\in\mathcal{J}_1(i)\coloneqq \mathcal{J}(i)\setminus\{1\}$ and $i\in\mathcal{I}$. In analogy to Subsection 4.3.2, we define We define

$$\bar{\eta}_n \;\coloneqq\; \max_{i\in\mathcal{I}}\,\max_{j\in\mathcal{J}_1(i)}\,\eta^n_{ij}\,,\quad \text{and}\quad \underline{\eta}_n \;\coloneqq\; \min_{i\in\mathcal{I}}\,\min_{j\in\mathcal{J}_1(i)}\,\eta^n_{ij}\,.$$

Lemma 5.4. *Consider a network with a dominant server pool, and assume $\varrho_n > 0$. Then for any $\theta \ge \theta^n_0 \coloneqq 2\,\frac{\max_i \mu^n_{i1}}{\min_i \mu^n_{i1}}$, and $\delta\in(0,1)$, there exist positive constants c_0 and c_1 such that (5.2) holds for all $n \ge n_0$.*

Proof. Suppose $\hat{x}\in\widetilde{\mathfrak{X}}^n$. A simple calculation using (5.27) shows that

$$\begin{aligned} \frac{1}{\theta}\,F^n_1(\hat{x},\hat{z}) \;=\; &\frac{\varrho_n}{m}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(-\hat{x}_i) + \sum_{i\in\mathcal{I}}\psi'_\varepsilon(-\hat{x}_i)\bigl(\hat{x}_i - u^c_i\langle e,\hat{x}\rangle^+\bigr) \\ &- \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}_1(i)} \psi'_\varepsilon(-\hat{x}_i)\bigl(\eta^n_{ij}-1\bigr)u^s_j\langle e,\hat{x}\rangle^- \\ &- \hat{\vartheta}^n\Biggl(\sum_{i\in\mathcal{I}}\psi'_\varepsilon(-\hat{x}_i)u^c_i + \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}_1(i)} \psi'_\varepsilon(-\hat{x}_i)\bigl(\eta^n_{ij}-1\bigr)u^s_j\Biggr)\,, \end{aligned} \tag{5.28}$$

and

$$\begin{aligned} F^n_2(\hat{x},\hat{z}) \;=\; &-\frac{\varrho_n}{m}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(\hat{x}_i) - \sum_{i\in\mathcal{I}}\psi'_\varepsilon(\hat{x}_i)\bigl(\hat{x}_i - u^c_i\langle e,\hat{x}\rangle^+\bigr) \\ &+ \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}_1(i)} \psi'_\varepsilon(\hat{x}_i)\bigl(\eta^n_{ij}-1\bigr)u^s_j\langle e,\hat{x}\rangle^- \\ &+ \hat{\vartheta}^n\Biggl(\sum_{i\in\mathcal{I}}\psi'_\varepsilon(\hat{x}_i)u^c_i + \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}_1(i)} \psi'_\varepsilon(\hat{x}_i)\bigl(\eta^n_{ij}-1\bigr)u^s_j\Biggr)\,. \end{aligned} \tag{5.29}$$

By (5.28) we obtain

$$\begin{aligned} \frac{1}{\theta}\,F^n_1(\hat{x},\hat{z}) \;&\le\; \varrho_n\varepsilon + \frac{m}{2} - \varepsilon\sum_{i\in\mathcal{I}}\hat{x}^-_i + \varepsilon\bigl(1-\underline{\eta}_n\bigr)^+\langle e,\hat{x}\rangle^- + \varepsilon\bigl(1-\underline{\eta}_n\bigr)^+\hat{\vartheta}^n \\ &\le\; \varrho_n\varepsilon + \frac{m}{2} - \varepsilon\sum_{i\in\mathcal{I}}\hat{x}^-_i + \varepsilon\bigl(1-\underline{\eta}_n\bigr)^+\Biggl(\sum_{i\in\mathcal{I}}(\hat{x}^-_i - \hat{x}^+_i) + \|\hat{x}^-\|_1\wedge\|\hat{x}^+\|_1\Biggr) \\ &\le\; \varrho_n\varepsilon + \frac{m}{2} - \varepsilon\bigl(\underline{\eta}_n\wedge 1\bigr)\|\hat{x}^-\|_1 \\ &\le\; \varrho_n\varepsilon + \frac{m}{2} - \frac{\varepsilon(1-\delta)}{2}\,\bigl(\underline{\eta}_n\wedge 1\bigr)\|\hat{x}\|_1 \quad \forall\,(\hat{x},\hat{z})\in\bigl(\widetilde{\mathfrak{X}}^n\setminus\mathcal{K}^+_\delta\bigr)\times\hat{\mathcal{Z}}^n(\hat{x})\,,\ \forall\,n\in\mathbb{N}\,, \end{aligned} \tag{5.30}$$

where we used (4.8) in the first inequality, Lemma 5.2 in the second, and (4.10) in the fourth.

Next, we estimate a bound for $F_2^n(\hat{x},\hat{z})$. Recall the definitions of $\mathcal{I}_\circ$, $\mathcal{J}_\circ$, $\mathcal{E}_\circ$, and $\mathcal{E}'_\circ$ in the proof of Lemma 5.2. Since $x \in \widetilde{\mathfrak{X}}^n$, we have $u_i^c = 0$ for all $i \in \mathcal{I}_\circ$, and $u_j^s = 0$ for all $j \in \mathcal{J}_\circ^c$. SWC policies are interpreted here as follows: customer classes must use all the servers in the leaf pools available to them before using servers in pool $j = 1$. Additionally, $\hat{x}_i \le -\sum_{(i,j)\in\mathcal{E}'_\circ} \bar{z}_{ij}$ for $i \in \mathcal{I}_\circ$, which implies that $\psi'_\varepsilon(x_i) = 0$ for all $i \in \mathcal{I}_\circ$ and $n > n_0$, by Notation 5.1. Hence, since $\sum_{i\in\mathcal{I}_\circ^c} \hat{x}_i > 0$, where $\mathcal{I}_\circ^c \equiv \mathcal{I}\setminus\mathcal{I}_\circ$, we have

$$\sum_{i\in\mathcal{I}} \psi'_\varepsilon(\hat{x}_i)\bigl(\hat{x}_i - u_i^c\langle e,\hat{x}\rangle^+\bigr) \;\ge\; \sum_{i\in\mathcal{I}_\circ^c} \psi'_\varepsilon(\hat{x}_i)\hat{x}_i - \varepsilon\sum_{i\in\mathcal{I}_\circ^c}\hat{x}_i - \varepsilon\sum_{i\in\mathcal{I}_\circ}\hat{x}_i \;\ge\; -\varepsilon\sum_{i\in\mathcal{I}_\circ}\hat{x}_i \tag{5.31}$$

by (4.12). Using (5.29) together with Remark 5.5 and (5.31), we obtain

$$\begin{aligned}
F_2^n(\hat{x},\hat{z}) \;&\le\; -\frac{\varrho_n\varepsilon}{m} + \varepsilon\hat{\vartheta}^n + \varepsilon\sum_{i\in\mathcal{I}_\circ}\hat{x}_i + \hat{\vartheta}^n\Biggl(\sum_{i\in\mathcal{I}_\circ}\sum_{j\in\mathcal{J}_1(i)} \psi'_\varepsilon(\hat{x}_i)\bigl(\eta_{ij}^n - 1\bigr)u_j^s\Biggr)\\
&\le\; -\frac{\varrho_n\varepsilon}{m} + \varepsilon(1-\varkappa_\circ^n)\sum_{i\in\mathcal{I}_\circ}\hat{x}_i\\
&\le\; \begin{cases} -\frac{\varrho_n\varepsilon}{m} - \frac{\varepsilon(1-\delta)}{2}\,(1-\varkappa_\circ^n)\,\widetilde{\varkappa}_\circ^n\,\|\hat{x}\|_1 & \text{for } (\hat{x},\hat{z}) \in \bigl(\widetilde{\mathfrak{X}}^n \cap (\mathcal{K}_0^+\setminus\mathcal{K}_\delta^+)\bigr)\times\hat{\mathcal{Z}}^n(\hat{x})\\ -\frac{\varrho_n\varepsilon}{2} & \text{for } (\hat{x},\hat{z}) \in \bigl(\widetilde{\mathfrak{X}}^n \cap \mathcal{K}_\delta^+\bigr)\times\hat{\mathcal{Z}}^n(\hat{x})\,,\end{cases}
\end{aligned} \tag{5.32}$$

for all $n \ge n_0$. Thus, the result follows by (5.30) and (5.32), noting also that the choice of θ implies that $V_1 \ge V_2^2$ on $\mathcal{K}_0^-$. □

5.1.3. *The diffusion-scale of networks with class-dependent service rates.* Recall from Subsection 4.3.2 that the drift in (5.4) reduces to

$$b^n(\hat{x},\hat{z}) \;=\; -\frac{\varrho_n}{m}B_1^n e - B_1^n\bigl(\hat{x} - \langle e,\hat{x}\rangle^+ u^c\bigr) + \hat{\vartheta}^n(\hat{x},\hat{z})B_1^n u^c\,. \tag{5.33}$$

where $B_1^n = \operatorname{diag}(\mu_1^n,\dots,\mu_m^n)$. Thus, the spare capacity ϱ_n is given by

$$\varrho_n \;=\; -\frac{1}{\sqrt{n}}\Biggl(\sum_{i\in\mathcal{I}}\frac{\lambda_i^n}{\mu_i^n} - \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}(i)}\xi_{ij}^* N_j^n\Biggr)\,.$$

Lemma 5.5. *Suppose that $\mu_{ij}^n = \mu_i^n$, for all $i \in \mathcal{I}$, and $\varrho_n > 0$. Then, for any $\theta \ge \theta_0^n := 2\,\frac{\mu^n_{\mathsf{max}}}{\mu^n_{\mathsf{min}}}$, and $\delta \in (0,1)$, the conclusions of Lemma 5.4 follow.*

Proof. Suppose $\hat{x} \in \widetilde{\mathfrak{X}}^n$. A simple calculation using (5.33) shows that

$$\frac{1}{\theta}\,F_1^n(\hat{x},\hat{z}) \;=\; \frac{\varrho_n}{m}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(-\hat{x}_i) + \sum_{i\in\mathcal{I}}\psi'_\varepsilon(-\hat{x}_i)\bigl(\hat{x}_i - u_i^c\langle e,\hat{x}\rangle^+\bigr) - \hat{\vartheta}^n\sum_{i\in\mathcal{I}}\psi'_\varepsilon(-\hat{x}_i)u_i^c\,, \tag{5.34}$$

$$F_2^n(\hat{x},\hat{z}) \;=\; -\frac{\varrho_n}{m}\sum_{i\in\mathcal{I}}\psi'_\varepsilon(\hat{x}_i) - \sum_{i\in\mathcal{I}}\psi'_\varepsilon(\hat{x}_i)\bigl(\hat{x}_i - u_i^c\langle e,\hat{x}\rangle^+\bigr) + \hat{\vartheta}^n\sum_{i\in\mathcal{I}}\psi'_\varepsilon(\hat{x}_i)u_i^c\,. \tag{5.35}$$

By (5.34), we obtain

$$\begin{aligned}
\frac{1}{\theta}\,F_1^n(\hat{x},\hat{z}) \;&\le\; \varrho_n\varepsilon + \frac{m}{2} - \varepsilon\|\hat{x}^-\|_1\\
&\le\; \varrho_n\varepsilon + \frac{m}{2} - \frac{\varepsilon(1-\delta)}{2}\|\hat{x}\|_1 \qquad \forall\,(\hat{x},\hat{z}) \in \bigl(\widetilde{\mathfrak{X}}^n\setminus\mathcal{K}_\delta^-\bigr)\times\hat{\mathcal{Z}}^n(\hat{x})\,,\ \forall\, n \ge n_0\,.
\end{aligned}$$

In computing the analogous bound to (5.32), there is a difference here. It is not the case here that $\psi'_\varepsilon(x_i) = 0$ for all $i \in \mathcal{I}_\circ$ and $n > n_0$.

So instead, recalling that $u^c_i = 0$ for all $i \in \mathcal{I}_\circ$, and since $\hat{x} \in \mathcal{K}_0^+$, we write

$$\begin{aligned}
-\sum_{i\in\mathcal{I}} \psi'_\varepsilon(\hat{x}_i)\bigl(\hat{x}_i - u^c_i\langle e,\hat{x}\rangle^+\bigr) + \hat{\vartheta}^n \sum_{i\in\mathcal{I}} \psi'_\varepsilon(\hat{x}_i) u^c_i \;&\le\; -\sum_{i\in\mathcal{I}} \psi'_\varepsilon(\hat{x}_i)\hat{x}_i + \varepsilon\langle e,\hat{x}\rangle + \varepsilon\hat{\vartheta}^n \\
&\le\; \varepsilon\Biggl(\hat{\vartheta}^n - \sum_{\ell\in\mathcal{I}_\circ} \hat{x}^-_\ell\Biggr) - \sum_{i\in\mathcal{I}_\circ} \psi'_\varepsilon(\hat{x}_i)\hat{x}_i \\
&\quad - \Biggl(\sum_{i\in\mathcal{I}^c_\circ} \psi'_\varepsilon(\hat{x}_i)\hat{x}_i - \varepsilon \sum_{i\in\mathcal{I}^c_\circ} \hat{x}_i\Biggr).
\end{aligned} \tag{5.36}$$

The third term on the right-hand side is nonpositive by (4.12). We also have

$$\hat{\vartheta}^n - \sum_{\ell\in\mathcal{I}_\circ} \hat{x}^-_\ell \;=\; -\sqrt{n} \sum_{(i,j)\in\mathcal{E}'_\circ} \bar{z}^n_{ij}\,, \tag{5.37}$$

and

$$-\sum_{i\in\mathcal{I}_\circ} \psi'_\varepsilon(\hat{x}_i)\hat{x}_i \;\le\; \sum_{\hat{x}^-_i \le \frac{1}{2m}\sqrt{n}\,\min_{i\sim j}\,\bar{z}^n_{ij}} \hat{x}^-_i \;\le\; \frac{1}{2}\sqrt{n}\,\min_{i\sim j}\,\bar{z}^n_{ij}\,. \tag{5.38}$$

Therefore, by (5.35), (5.36)–(5.38) and Remark 5.5, we obtain

$$\begin{aligned}
F^n_2(\hat{x},\hat{z}) \;&\le\; \frac{\varepsilon}{2}\Biggl(\hat{\vartheta}^n - \sum_{\ell\in\mathcal{I}_\circ} \hat{x}^-_\ell\Biggr) \\
&\le\; \frac{\varepsilon}{2}(1-\varkappa^n_\circ)\,\widetilde{\varkappa}^n_\circ\,\|\hat{x}^-\|_1 \qquad \forall\,(\hat{x},\hat{z}) \in \bigl(\widetilde{\mathfrak{X}}^n \cap \mathcal{K}_0^+\bigr) \times \hat{\mathcal{Z}}^n(\hat{x})\,.
\end{aligned}$$

The rest follows as in Lemma 5.4. □

6. Concluding Remarks

In this paper we have identified a large family of networks for which the uniform exponential ergodic property is established via a unified approach using a common Lyapunov function. Many interesting but challenging questions remain to be studied in future work. It is important to identify other types of networks that possess the uniform exponential ergodicity. For example, it is shown in [29, Section 4] that for networks with pool-dependent service rates, the invariant measures of the diffusion-scaled processes are tight under a specific (LQAFS-LB) scheduling policy. It will be interesting to investigate if this class of networks possess the uniform exponential ergodicity property. Observe that the dominating matrix of drift is not diagonal for these models. Thus, the Lyapunov functions used in this paper can no longer be used, and it requires to construct a new type of Lyapunov function to treat both the limiting diffusion and the prelimit diffusion-scaled processes.

Acknowledgment

This work is supported in part by the Army Research Office through grant W911NF-17-1-0019, and in part by NSF grants DMS-1715210, CMMI-1635410, and DMS/CMMI-1715875, and in part by the Office of Naval Research through grant N00014-16-1-2956 and was approved for public release under DCN #43-5454-19.

References

[1] Z. Aksin, M. Armony, and V. Mehrotra, *The modern call center: a multi-disciplinary perspective on operations management research*, Production and Operations Management **16** (2007), 665–688.

[2] A. Arapostathis, A. Biswas, and G. Pang, *Ergodic control of multi-class $M/M/N+M$ queues in the Halfin-Whitt regime*, Ann. Appl. Probab. **25** (2015), no. 6, 3511–3570. MR3404643

[3] A. Arapostathis, V. S. Borkar, and M. K. Ghosh, *Ergodic control of diffusion processes*, Encyclopedia of Mathematics and its Applications, vol. 143, Cambridge University Press, Cambridge, 2012. MR2884272

[4] A. Arapostathis, H. Hmedi, and G. Pang, *On uniform exponential ergodicity of Markovian multiclass many-server queues in the Halfin–Whitt regime*, Math. Oper. Res. (2020). to appear.

[5] A. Arapostathis and G. Pang, *Ergodic diffusion control of multiclass multi-pool networks in the Halfin-Whitt regime*, Ann. Appl. Probab. **26** (2016), no. 5, 3110–3153. MR3563203

[6] A. Arapostathis and G. Pang, *Infinite-horizon average optimality of the N-network in the Halfin-Whitt regime*, Math. Oper. Res. **43** (2018), no. 3, 838–866. MR3846075

[7] A. Arapostathis and G. Pang, *Infinite horizon asymptotic average optimality for large-scale parallel server networks*, Stochastic Process. Appl. **129** (2019), no. 1, 283–322. MR3906999

[8] A. Arapostathis, G. Pang, and N. Sandrić, *Ergodicity of a Lévy-driven SDE arising from multiclass many-server queues*, Ann. Appl. Probab. **29** (2019), no. 2, 1070–1126. MR3910024

[9] M. Armony, S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, and G. B. Yom-Tov, *On patient flow in hospitals: a data-based queueing-science perspective*, Stoch. Syst. **5** (2015), no. 1, 146–194. MR3442392

[10] R. Atar, *A diffusion model of scheduling control in queueing systems with many servers*, Ann. Appl. Probab. **15** (2005), no. 1B, 820–852. MR2114991

[11] R. Atar, *Scheduling control for queueing systems with many servers: asymptotic optimality in heavy traffic*, Ann. Appl. Probab. **15** (2005), no. 4, 2606–2650. MR2187306

[12] S. Borst, A. Mandelbaum, and M. I. Reiman, *Dimensioning large call centers*, Oper. Res. **52** (2004), no. 1, 17–34. MR2066238

[13] A. Braverman, J. G. Dai, and M. Miyazawa, *Heavy traffic approximation for the stationary distribution of a generalized Jackson network: the BAR approach*, Stoch. Syst. **7** (2017), no. 1, 143–196. MR3663340

[14] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, *Statistical analysis of a telephone call center: a queueing-science perspective*, J. Amer. Statist. Assoc. **100** (2005), no. 469. MR2166068

[15] A. Budhiraja and C. Lee, *Stationary distribution convergence for generalized Jackson networks in heavy traffic*, Math. Oper. Res. **34** (2009), no. 1, 45–56. MR2542988

[16] A. B. Dieker and X. Gao, *Positive recurrence of piecewise Ornstein-Uhlenbeck processes and common quadratic Lyapunov functions*, Ann. Appl. Probab. **23** (2013), no. 4, 1291–1317. MR3098433

[17] D. Gamarnik and A. L. Stolyar, *Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime: asymptotics of the stationary distribution*, Queueing Syst. **71** (2012), no. 1-2, 25–51. MR2925789

[18] D. Gamarnik and A. Zeevi, *Validity of heavy traffic steady-state approximation in generalized Jackson networks*, Ann. Appl. Probab. **16** (2006), no. 1, 56–90. MR2209336

[19] N. Gans, G. Koole, and A. Mandelbaum, *Telephone call centers: Tutorial, review and research prospects*, Manufacturing and Service Operations Management **5** (2003), 79–141.

[20] O. Garnett, A. Mandelbaum, and M. I. Reiman, *Designing a call center with impatient customers*, Manufacturing and Service Operations Management **4** (2002), no. 3, 208–227.

[21] I. Gurvich, *Validity of heavy-traffic steady-state approximations in multiclass queueing networks: the case of queue-ratio disciplines*, Math. Oper. Res. **39** (2014), no. 1, 121–162. MR3173006

[22] I. Gyöngy and N. Krylov, *Existence of strong solutions for Itô's stochastic equations via approximations*, Probab. Theory Related Fields **105** (1996), no. 2, 143–158. MR1392450

[23] S. Halfin and W. Whitt, *Heavy-traffic limits for queues with many exponential servers*, Oper. Res. **29** (1981), no. 3, 567–588. MR629195

[24] H. Hmedi, A. Arapostathis, and G. Pang, *On system-wide safety staffing of large-scale parallel server networks*, Operations Research (2022).

[25] P. Shi, M. Chou, J. G. Dai, D. Ding, and J. Sim, *Models and insights for hospital inpatient operations: Time-dependent ed boarding time*, Management Science **62** (2016), no. 1, 1–28.

[26] A. L. Stolyar, *Diffusion-scale tightness of invariant distributions of a large-scale flexible service system*, Adv. in Appl. Probab. **47** (2015), no. 1, 251–269. MR3327324

[27] A. L. Stolyar, *Tightness of stationary distributions of a flexible-server system in the Halfin-Whitt asymptotic regime*, Stoch. Syst. **5** (2015), no. 2, 239–267. MR3442428

[28] A. L. Stolyar and E. Yudovina, *Tightness of invariant distributions of a large-scale flexible service system under a priority discipline*, Stoch. Syst. **2** (2012), no. 2, 381–408. MR3354771

[29] A. L. Stolyar and E. Yudovina, *Systems with large flexible server pools: instability of "natural" load balancing*, Ann. Appl. Probab. **23** (2013), no. 5, 2099–2138. MR3134731

[30] M. van der Boor, S. C Borst, J. S. van Leeuwaarden, and D. Mukherjee, *Scalable load balancing in networked systems: A survey of recent advances*, ArXiv e-prints **1806.05444** (2018), available at https://arxiv.org/abs/1806.05444.

[31] J. S. Van Leeuwaarden, B. W. Mathijsen, and B. Zwart, *Economies-of-scale in resource sharing systems: tutorial and partial review of the QED heavy-traffic regime*, ArXiv e-prints **1706.05397** (2017), available at https://arxiv.org/abs/1706.05397.

[32] W. Whitt, *Understanding the efficiency of multi-server service systems*, Management Science **38** (1992), no. 5, 708–723.

[33] R. J. Williams, *On dynamic scheduling of a parallel server system with complete resource pooling*, Fields Inst. Commun., vol. 28, Amer. Math. Soc., Providence, RI, 2000. MR1788708

[34] H.-Q. Ye and D. D. Yao, *Diffusion limit of fair resource control—stationarity and interchange of limits*, Math. Oper. Res. **41** (2016), no. 4, 1161–1207. MR3544792

[35] H.-Q. Ye and D. D. Yao, *Justifying diffusion approximations for multiclass queueing networks under a moment condition*, Ann. Appl. Probab. **28** (2018), no. 6, 3652–3697. MR3861823