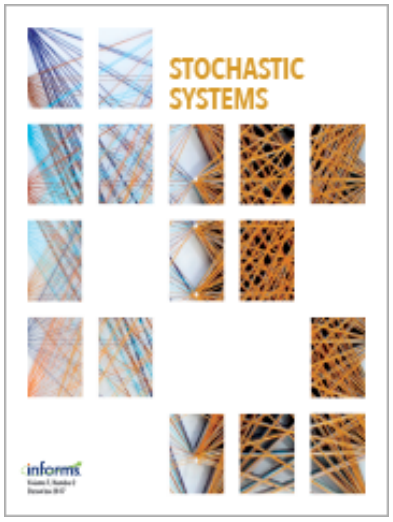
This article was downloaded by: [216.73.217.172] On: 15 November 2026, At: 06:06
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA



Stochastic Systems

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

Tightness of Stationary Distributions of a Flexible-Server System in the Halfin-Whitt Asymptotic Regime

Alexander L. Stolyar

To cite this article:
Alexander L. Stolyar (2015) Tightness of Stationary Distributions of a Flexible-Server System in the Halfin-Whitt Asymptotic Regime. Stochastic Systems 5(2):239-267. https://doi.org/10.1287/14-SSY139

This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as "*Stochastic Systems*. Copyright 2015 The Author(s). https://doi.org/10.1287/14-SSY139, used under a Creative Commons Attribution License: http://creativecommons.org/licenses/by/4.0/."

Copyright © 2015, The author(s)

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

Stochastic Systems
2015, Vol. 5, No. 2, 239–267
DOI: 10.1214/14-SSY139

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

TIGHTNESS OF STATIONARY DISTRIBUTIONS OF A FLEXIBLE-SERVER SYSTEM IN THE HALFIN-WHITT ASYMPTOTIC REGIME

By Alexander L. Stolyar

Lehigh University

A large-scale flexible service system with two large server pools and two types of customers is considered. Servers in pool 1 can only serve type 1 customers, while server in pool 2 are flexible – they can serve both types 1 and 2. (This is a so-called "N-system.") The customer arrival processes are Poisson and customer service requirements are independent exponentially distributed. The service rate of a customer depends both on its type and the pool where it is served. A priority service discipline, where type 2 has priority in pool 2, and type 1 prefers pool 1, is considered. We consider the Halfin-Whitt asymptotic regime, where the arrival rate of customers and the number of servers in each pool increase to infinity in proportion to a scaling parameter n, while the overall system capacity exceeds its load by $O(\sqrt{n})$.

For this system we prove tightness of diffusion-scaled stationary distributions. Our approach relies on a single common Lyapunov function $G^{(n)}(x)$, depending on parameter n and defined on the entire state space as a functional of the *drift-based fluid limits* (DFL). Specifically, $G^{(n)}(x) = \int_0^\infty g(y^{(n)}(t))dt$, where $y^{(n)}(\cdot)$ is the DFL starting at x, and $g(\cdot)$ is a "distance" to the origin. ($g(\cdot)$ is same for all n). The key part of the analysis is the study of the (first and second) derivatives of the DFLs and function $G^{(n)}(x)$. The approach, as well as many parts of the analysis, are quite generic and may be of independent interest.

1. Introduction. In this paper we consider a large-scale service system in the so-called Halfin-Whitt asymptotic regime. Such systems received a lot of attention in the literature, especially in the past 10-15 year, because they find a variety of applications, including, e.g., large customer contact centers [1, 11] and large computer farms in network clouds. The Halfin-Whitt regime, introduced originally in [14], is such that the system capacity (roughly, number of servers) increases in proportion to a scaling parameter n, and exceeds the system load by $O(\sqrt{n})$. It is attractive because it al-

Received March 2014.
AMS 2000 subject classifications: Primary 60K25, 60F17
Keywords and phrases: Many server models, drift-based fluid limit, diffusion limit, tightness of stationary distributions, limit interchange, common Lyapunov function

lows – in principle, under a good control algorithm – to achieve both good performance (e.g. waiting times) and high resource utilization.

In the Halfin-Whitt regime, the stochastic process describing the system behavior is usually studied under *diffusion scaling*, i.e. it is centered at the system equilibrium point and scaled down by $n^{-1/2}$. This name reflects the fact that, in the limit on $n \to \infty$, on any finite time interval, the sequence of diffusion-scaled processes $Y^{(n)}(\cdot)$ "typically" converges to a diffusion process $Y(\cdot)$. Then, a fundamental question is whether or not the following *limit interchange* property holds: *the limit of stationary distributions of* $Y^{(n)}(\cdot)$ *is equal to the stationary distribution of* $Y(\cdot)$. In turn, the key difficulty in establishing the limit interchange property, is verifying the *(stationary distribution) tightness* property: *the family of stationary distributions of* $Y^{(n)}(\cdot)$ *is tight*.

The tightness property in the Halfin-Whitt regime is usually difficult to verify even for systems with single pool of homogeneous servers, if there is more than one type of arriving customers and/or the service time distribution is non-exponential; see [3, 7–9] for the results in this direction. (We note that the problems of verifying the tightness and limit interchange exists not only in the Halfin-Whitt regime, but also in the so-called *conventional heavy traffic* regime; see [2, 10, 12].) More general models, where there are multiple flexible server pools with different capabilities (service rates) w.r.t. different customer types, pose additional challenges. The key additional difficulty is that for such systems the state space is "fractured" into multiple domains, where the process dynamics is very different. Papers [16–18] contain tightness / limit interchange results for some flexible multi-pool models; although, [16, 17] consider a strictly subcritical load regime (different from Halfin-Whitt), in which the capacity exceeds the load by $O(n)$.

One approach for verifying the stationary distribution tightness is to find a single common Lyapunov function, for which an appropriate "negative expected drift" condition can be established. This approach is used in [3, 8, 9, 18]. (Papers [7, 16, 17] use different approaches, not relying on a single Lyapunov function.) Of course, finding/constructing a suitable Lyapunov function is usually the key challenge. For example, paper [3], which proves tightness for a single-pool model with first-come-first-serve discipline and phase-type service time distribution, uses an elaborate *common quadratic Lyapunov function*, of the type proposed in [4]; the tightness result in [3] also requires that the customers waiting in the queue abandon at positive rate. And again, finding single common Lyapunov function is further complicated for flexible multi-pool systems.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

1.1. *Paper contributions.* We consider a flexible multi-pool system with two customer types and two server pools (the so-called "N-system"), under Markovian assumptions (the customer arrival processes are Poisson and customer service requirements are independent exponentially distributed) and under a priority discipline. The service rate of a customer depends both on its type and the pool where it is served. We study the Halfin-Whitt asymptotic regime, and prove the stationary distribution tightness result, Theorem 2, which implies the limit-interchange, Corollary 4. (These results extend to a more general class of Markovian systems under priority disciplines, as discussed in Section 6.)

The state space of the diffusion-scaled process for N-system has five domains, where the process drift is given by different affine functions; the domain boundaries depend on parameter n. Nevertheless, we construct a single Lyapunov function $G^{(n)}(x)$ (depending on parameter n) on the entire state space, as a functional of the *drift-based fluid limits* (DFL), which are the deterministic trajectories defined by the drift of the process. Specifically,

$$G^{(n)}(x) = \int_0^\infty g(y^{(n)}(t))dt, \tag{1}$$

where $y^{(n)}(\cdot)$ is the DFL starting at x, and $g(\cdot)$ is a "distance" to the origin. (Function $g(\cdot)$ does *not* depend on n.) For a Lyapunov function of this type, in a setting more general than needed for the proof of Theorem 2, we give sufficient conditions for the tightness in Theorem 5; the key condition is a (uniform in n) bound on the Lyapunov function second derivatives. This result may be of independent interest.

The proof of Theorem 2 verifies the conditions of Theorem 5 for the N-system. This requires the analysis of the DFL structure, and of the (first and second) derivatives of DFLs and corresponding functionals $G^{(n)}(x)$ on the initial state x; it also requires an appropriate choice of the "distance" $g(\cdot)$. Many parts of this analysis are quite generic and may also be of independent interest.

As will be illustrated below in Section 2, for a *deterministic* dynamic system, with trajectories $y^{(n)}(\cdot)$ defined by a continuous derivative-field, the function $G^{(n)}(x)$ given by (1) is a natural Lyapunov function, as long as it is well defined (the integral in (1) is finite). In queueing networks literature, this observation is used, for example, in [15, 19] to establish the existence of a Lyapunov function for stable *deterministic* fluid models. This observation, however, does *not* imply that $G^{(n)}(x)$ defined by (1) via DFLs $y^{(n)}(\cdot)$ can serve as a Lyapunov function for a (family of) *random* process(es). In this paper we give sufficient conditions under which Lyapunov functions $G^{(n)}(x)$

can be used to establish tightness of stationary distributions, and then verify these conditions for the N-system.

A Lyapunov function similar in spirit to (1) was used in [5] to establish a sufficient condition for positive recurrence of a semimartingale reflecting Brownian motion in the positive orthant. (In [5], the solutions to the Skorohod problem, for the trajectories determined by the process drift alone, are the DFLs in our terminology.) Obtaining the Lyapunov function second derivative bounds is also a key part of the analysis in [5]. We note, however, that our basic model, the problem, the structure of the (family of) process(es) and corresponding DFLs, the form of function $g(\cdot)$, and the analysis of the Lyapunov function derivatives are completely different.

1.2. *Layout of the rest of the paper.* In Section 2, we informally discuss our general approach and the Lyapunov function construction. Section 3 formally defines the N-system, the Halfin-Whitt regime for it, and states the tightness (Theorem 2) and the limit-interchange (Corollary 4) results. In Section 4, in a setting more general than needed for the N-system, we give a formal construction of the DFLs and the Lyapunov function, and sufficient conditions for the tightness (Theorem 5). Section 5 contains the proof of Theorem 2; here we choose a specific "distance" function g and verify the conditions of Theorem 5 for the N-system. A generalization of the N-system, for which our results still hold, is described in Section 6. Finally, in Section 7, we discuss our approach and results.

1.3. *Basic notation.* Symbols $\mathbb{R}, \mathbb{R}_+, \mathbb{Z}, \mathbb{Z}_+$ denote the sets of real, real non-negative, integer, and integer non-negative numbers, respectively. In the Euclidean space $\mathbb{R}^I$ (of dimension $I \geq 1$): $|x|$ denotes standard Euclidean norm of vector $x = (x_1, \ldots, x_I)$, while $\|x\| = \sum_i |x_i|$ denotes its L_1-norm; scalar product of two vectors is denoted $x \cdot y = \sum_i x_i y_i$; $diag(x)$ denotes diagonal square matrix, with diagonal elements given by x; we write simply 0 for a zero matrix or vector; vectors are written as row-vectors, but in matrix expressions they are viewed as column-vectors (without using a transposition sign). For real numbers u and w: $u \vee w = \max\{u, w\}$, $u \wedge w = \min\{u, w\}$, and $\lfloor u \rfloor$ denotes the largest integer not greater than u.

For a vector-function $y(\cdot) = (y(t),\ t \geq 0)$, we denote $\|y(\cdot)\| = \sup_{[0,\infty)} \|y(t)\|$. Abbreviation *u.o.c.* means *uniform on compact sets* convergence. If $X(t),\ t \geq 0$, is a Markov process, we write $X(\infty)$ for a random element with the distribution equal to a stationary distribution of the process. (In all cases considered in this paper, the stationary distribution will be unique.) Symbol $\Rightarrow$ denotes convergence in distribution of random elements; random processes are random elements in the appropriate Skorohod space.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

For a condition/event H, the indicator function $I\{H\}$ is equal to 1 when H holds and 0 otherwise.

2. The intuition for the Lyapunov function construction. The discussion in this entire section is informal. Consider a deterministic dynamic system governed by ODE

$$(d/dt)y = v(y), \tag{2}$$

where state y is a vector, and the vector-field $v(\cdot)$ is Lipschitz continuous. Suppose the system has unique stable point 0. Let $g(x)$ be a non-negative continuous (and sufficiently smooth) function, which measures a "distance" from 0. (In our results, we will use $g(x)$ which is a smooth approximation of L_1-norm $\|x\|$.) Suppose that for any initial state $y(0) = x$ the trajectory $y(t),\ t \geq 0$ converges to 0 and, moreover,

$$G(x) = \int_0^\infty g(y(t))dt < \infty. \tag{3}$$

Then $G(\cdot)$ is a Lyapunov function for this dynamic system, in the sense that

$$(d/dt)G(y(t)) = G'(y(t)) \cdot v(y(t)) = -g(y(t)),$$

where G' denotes the gradient of G. (The first equality is immediate from (2), and $(d/dt)G(y(t)) = -g(y(t))$ is from $G(y(t)) = \int_t^\infty g(y(s))ds$.)

Suppose now that instead of a deterministic system we have a Markov process $Y(\cdot)$, for which vector-field $v(\cdot)$ gives the drift. Then we can define deterministic trajectories $y(\cdot)$, and function $G(\cdot)$, the same way as above. (The trajectories $y(\cdot)$ we call drift-based fluid limits (DFL).) Suppose further that the process generator A is such that

$$AG(y) = G'(y) \cdot v(y) + H(y), \qquad |H(y)| \leq C_0\|G''(y)\|, \tag{4}$$

where C_0 is a constant and G'' denotes the Hessian matrix of second derivatives. (To interpret (4) one can think, for example, of a diffusion process with bounded diffusion coefficients. In this paper we will work *not* with diffusion processes, but rather with diffusion-scaled processes for our queueing system – their behavior can be very different from that of diffusions, especially when the system state is "far" from the equilibrium point. Nevertheless, the process generator will have form (4).) Then, we have

$$AG(y) \leq G'(y) \cdot v(y) + C_0\|G''(y)\| = -g(y) + C_0\|G''(y)\|.$$

If we can show that

$$\|G''(y)\| \le C_1 g(y) + C_2 \tag{5}$$

with a sufficiently small C_1, then for some $\epsilon > 0$ and $\kappa > 0$,

$$AG(y) \le -\epsilon g(y) + \kappa. \tag{6}$$

This is a Lyapunov-Foster type condition from which we can obtain the steady-state bound $\mathbb{E}g(Y(\infty)) \le \kappa/\epsilon$, where $Y(\infty)$ is $Y(t)$ when the process is in stationary regime.

Finally, suppose we consider a family of processes $Y(\cdot) = Y^{(n)}(\cdot)$, with the drift $v(\cdot)$ and generator A depending on n. If for some common function g such that $g(x) \to \infty$ as $\|x\| \to \infty$, we can derive estimates (4)–(6) with constants independent of n, then $\mathbb{E}g(Y^{(n)}(\infty))$ is bounded uniformly in n, and therefore the family of stationary distributions of $Y^{(n)}(\cdot)$ is tight.

This is the program that we implement in this paper, for the sequence of diffusion-scaled processes for the N-system. The difficult part is obtaining the second derivative bound (5). Since G is defined as a functional of the DFLs $y(\cdot)$, this involves the analysis of the dependence of DFLs on the initial state.

3. *N*-system with absolute priority. Consider a so-called N-system, with absolute priorities. (See Fig. 1.) There are two customer types, arriving according to independent Poisson processes with rates $\Lambda_1 > 0$ and $\Lambda_2 > 0$, respectively. There are two server pools, with B_1 and B_2 identical servers, respectively. The total service requirement of any customer is an independent, exponentially distributed random variable with mean 1. A customer of type 2 can only be served by a server in pool 2, and if it does receive service, it does so at rate $\mu_{22} > 0$. A customers of type 1 can be served by a server in either pool 1 or 2, with service rates being $\mu_{11} > 0$ and $\mu_{12} > 0$, respectively. Type 2 customers have absolute (preemptive) priority (in pool 2); namely, if there are X_2 type 2 customers in the system, as many of them as possible, $X_2 \wedge B_2$, receive service in pool 2, and the remaining $X_2 - X_2 \wedge B_2 = (X_2 - B_2) \vee 0$ wait in the queue. (Here $\wedge$ and $\vee$ denote minimum and maximum, respectively.) Therefore, the total service rate of all type 2 customers is

$$\mu_{22}(X_2 \wedge B_2). \tag{7}$$

The type 1 customers have absolute preference to be served in pool 1, and have lower preempt-resume priority in pool 2. Namely, if there are X_1 type 1

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

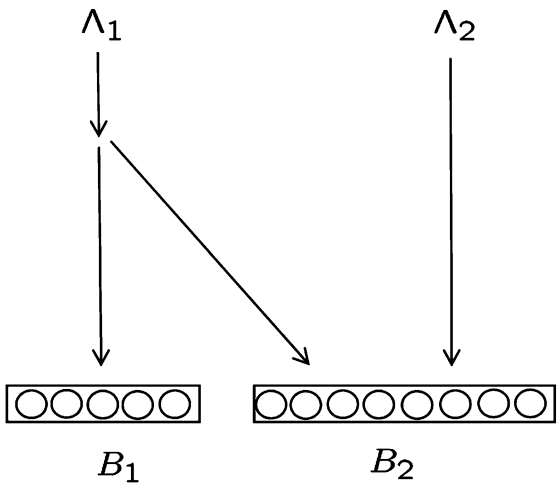


FIG 1. *N-system.*

customers in the system, then $X_1 \wedge B_1$ of them are served in pool 1, $[(X_1 - B_1) \vee 0] \wedge [(B_2 - X_2) \vee 0]$ are served in pool 2, and the remaining $[X_1 - (B_1 + B_2) + (X_2 \wedge B_2)] \vee 0$ wait in queue. The total service rate of all type 1 customers is

$$\mu_{11}\{X_1 \wedge B_1\} + \mu_{12}\{[(X_1 - B_1) \vee 0] \wedge [(B_2 - X_2) \vee 0]\}. \tag{8}$$

We consider a sequence of such systems, indexed by a positive scaling parameter n, increasing to infinity. (See Fig. 2.) In a system with parameter n,

$$\Lambda_1 = \lambda_1 n, \quad \Lambda_2 = \lambda_2 n, \tag{9}$$

$$B_1 = \psi_{11} n, \quad B_2 = \psi_{12} n + \psi_{22} n + b\sqrt{n}, \tag{10}$$

where the positive parameters $b, \lambda_1, \lambda_2, \psi_{11}, \psi_{12}, \psi_{22}$ are such that

$$\lambda_2 = \psi_{22}\mu_{22}, \quad \lambda_1 = \psi_{11}\mu_{11} + \psi_{12}\mu_{12}. \tag{11}$$

Given this definition, and the priorities, the system "desired operating point," which we will refer to as *equilibrium point*, is such that $X_2 = \psi_{22} n$ and $X_1 = \psi_{11} n + \psi_{12} n$, where type 1 customers occupy the entire pool 1 and $\psi_{12} n$ servers in pool 2; the equilibrium point is such that $b\sqrt{n}$ servers in pool 2 are idle – this is the "margin" by which system capacity exceeds its load. (Again, see Fig. 2.)

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

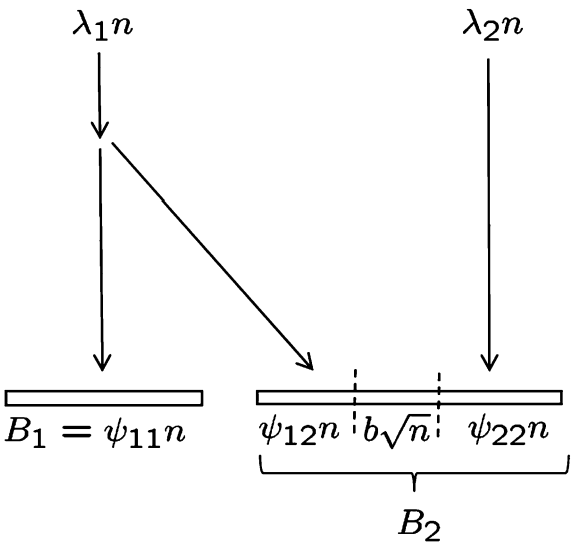


FIG 2. *N-system in Halfin-Whitt asymptotic regime.*

REMARK 1. To be precise, in the definition of the sequence of systems, we need to make sure that B_1 and B_2 are integer. Equations (10), as written above, assume that B_1 and B_2 "happen to be" integer. We make this assumption throughout the paper to simplify the exposition, while maintaining rigor of the results and arguments. More specifically, we could replace (10) with, for example,

$$B_1 = \lfloor \psi_{11} n \rfloor, \quad B_2 = \lfloor \psi_{12} n + \psi_{22} n + b\sqrt{n} \rfloor. \tag{12}$$

If we do that, it is easy to check that for each n we can choose numbers $\psi_{ij}^{(n)}$, $(ij) = (11), (12), (22)$, and $b^{(n)}$, such that: $|\psi_{ij}^{(n)} - \psi_{ij}| \leq \kappa/n$ and $|b^{(n)} - b| \leq \kappa/\sqrt{n}$ for some constant $\kappa > 0$; (12) can be rewritten as

$$B_1 = \psi_{11}^{(n)} n, \quad B_2 = \psi_{12}^{(n)} n + \psi_{22}^{(n)} n + b^{(n)}\sqrt{n}; \tag{13}$$

and (11) can be rewritten as

$$\lambda_2 = \psi_{22}^{(n)} \mu_{22}, \quad \lambda_1 = \psi_{11}^{(n)} \mu_{11} + \psi_{12}^{(n)} \mu_{12}. \tag{14}$$

The sequence of systems will then be defined by (9), (13), (14). Then, the entire analysis in this paper will hold as is, with ψ_{ij} and b replaced everywhere with $\psi_{ij}^{(n)}$ and $b^{(n)}$, respectively. (We note that the components of the equilibrium point, namely $X_2 = \psi_{22}^{(n)} n$ and $X_1 = \psi_{11}^{(n)} n + \psi_{12}^{(n)} n$, need *not* be integer.)

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

It is easy to see that for each n the process $X^{(n)}(t) = (X_1^{(n)}(t), X_2^{(n)}(t))$, $t \geq 0$, is continuous-time countable irreducible Markov chain, with the state space being (for each n) $\mathbb{Z}_+^2$. Further, it is not difficult to check that, for each sufficiently large n, this Markov process is positive recurrent, and therefore has unique stationary distribution. Indeed, due to absolute priority, type 2 customers "do not see" type 1, and therefore $X_2^{(n)}(\cdot)$ in itself is a positive recurrent Markov chain, which in steady-state occupies on average $\psi_{22}n$ servers in pool 2. This means that on average $\psi_{12}n + b\sqrt{n}$ servers in pool 2 are available to serve type 1 customers; this is in addition to all $\psi_{11}n$ servers in pool 1 which are available exclusively to type 1; therefore, the average total service capacity available to type 1 is

$$\psi_{11}n\mu_{11} + (\psi_{12}n + b\sqrt{n})\mu_{12} = \lambda_1 n + b\mu_{12}\sqrt{n} > \lambda_1 n.$$

More details of a positive recurrence proof are given in Appendix A.

The diffusion-scaled version $\hat{X}^{(n)}(t) = (\hat{X}_1^{(n)}(t), \hat{X}_2^{(n)}(t))$ of the process $X^{(n)}(t)$ is defined by centering at the equilibrium point and rescaling by $1/\sqrt{n}$:

$$\hat{X}_1^{(n)} = (X_1^{(n)} - \psi_{11}n - \psi_{12}n)/\sqrt{n}, \quad \hat{X}_2^{(n)} = (X_2^{(n)} - \psi_{22}n)/\sqrt{n}. \tag{15}$$

THEOREM 2. *For some $C > 0$ and all sufficiently large n,*

$$\mathbb{E}\|\hat{X}^{(n)}(\infty)\| \leq C.$$

The proof of Theorem 2 is given in the rest of this paper. It relies on a Lyapunov function (depending on n), being a functional of a fluid trajectory, determined by the process drift. Such fluid trajectories will be referred to as drift-based fluid limits (DFL). In the rest of this section we define DFLs for the N-system under consideration, and give motivation for the form of Lyapunov function. Then, in Section 4, we give the Lyapunov function construction and sufficient tightness conditions (Theorem 5) in a setting that is more general than needed for the N-system. In the following sections we verify the conditions of Theorem 5 for the N-system, thus proving Theorem 2.

For each n, for the unscaled process $X^{(n)}(\cdot)$, we define a drift function (vector field) $V^{(n)} = (V_1^{(n)}, V_2^{(n)})$ for $x = (x_1, x_2) \in \mathbb{R}_+^2$. (Note that it is defined on $\mathbb{R}_+^2$, and not just on the lattice $\mathbb{Z}_+^2$.) It is defined in the natural way, as the difference of arrival and service rates (see (7)–(8)):

$$V_1^{(n)} = V_1^{(n)}(x) = \Lambda_1 - \mu_{11}\{x_1 \wedge B_1\} - \mu_{12}\{[(x_1 - B_1) \vee 0] \wedge [(B_2 - x_2) \vee 0]\}, \tag{16}$$

$$V_2^{(n)} = V_2^{(n)}(x) = \Lambda_2 - \mu_{22}(x_2 \wedge B_2), \tag{17}$$

where $\Lambda_1, \Lambda_2, B_1, B_2$ are the functions of n given in (9)–(11).

Let us denote by L_n the affine mapping $X^{(n)} \to \hat{X}^{(n)}$, defined by (15). Then, the state space of $\hat{X}^{(n)}$ is $\mathcal{S}^{(n)} \equiv L_n\mathbb{Z}_+^2 \subset \mathcal{X}^{(n)} \equiv L_n\mathbb{R}_+^2 \subset \mathbb{R}^2$. Specifically, $\mathcal{X}^{(n)} = \{x \mid x_1 \geq -\psi_{11}\sqrt{n} - \psi_{12}\sqrt{n},\ x_2 \geq -\psi_{22}\sqrt{n}\}$. The drift function for $\hat{X}^{(n)}$ is defined accordingly:

$$v^{(n)}(x) = (1/\sqrt{n})V^{(n)}(L_n^{-1}x), \quad x \in \mathcal{X}^{(n)}.$$

We emphasize that $v^{(n)}(x)$ is defined on the continuous convex set $\mathcal{X}^{(n)}$, which contains the discrete state space $\mathcal{S}^{(n)}$. It is important, however, that *at each point $x \in \mathcal{S}^{(n)}$, $v^{(n)}(x)$ gives exactly the average drift of the process.* Namely,

$$v^{(n)}(x) = \sum_{x'} (x' - x)\nu^{(n)}(x, x'), \tag{18}$$

where $\nu^{(n)}(x, x')$ is the Markov process transition rate from state x to state x'; note that there is only a finite number of "neighbor" states x' for which $\nu^{(n)}(x, x') > 0$.

As $n \to \infty$, set $\mathcal{X}^{(n)}$ monotonically increases and converges to $\mathbb{R}^2$.

It is easy to observe that $v^{(n)}(x) = 0$ if and only if $x = 0$; also, uniformly in n, $v^{(n)}(x)$ is Lipschitz continuous. Given Lipschitz continuity of $v^{(n)}$, for any $x \in \mathcal{X}^{(n)}$ there is a unique solution $y^{(n)}(t),\ t \geq 0$, to the ODE

$$(d/dt)y^{(n)}(t) = v^{(n)}(y^{(n)}(t)), \quad y^{(n)}(0) = x.$$

The solution stays within $\mathcal{X}^{(n)}$ for all $t \geq 0$. (Indeed, for each i, $v_i^{(n)}(x) = \lambda_i\sqrt{n}$ when x_i is at its lower bound – see the definition of $\mathcal{X}^{(n)}$.) This trajectory $y^{(n)}(t),\ t \geq 0$, will be called the *drift-based fluid limit* (DFL), starting from x.

As we will show later in Section 5.1, each DFL is such that $y^{(n)}(t) \to 0$ as $t \to \infty$. Moreover, after a finite time this convergence is exponentially fast, so that

$$\int_0^\infty \|y^{(n)}(t)\|dt < \infty.$$

The Lyapunov function we will use to prove Theorem 2 is

$$G^{(n)}(x) = \int_0^\infty g(y^{(n)}(t))dt < \infty,$$

where $y^{(n)}(\cdot)$ is the DFL starting from x, and $g(\cdot)$ is a smooth non-negative function (common for all n) approximating $\|\cdot\|$.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

REMARK 3. In the literature on the steady-state tightness in the Halfin-Whitt asymptotic regime, deterministic trajectories defined by the drift vector field, which we call DFLs, are considered in e.g. [3], where they are called fluid models. However, the way we use DFLs in this paper – namely, to directly construct a Lyapunov function from them – is completely different from their use in [3].

3.1. *Limit interchange.* We conclude this section by noting that the tightness of stationary distributions of the processes $\hat{X}^{(n)}(\cdot)$, which follows from Theorem 2, allows us to easily establish the limit interchange result, given in Corollary 4 below.

Recall that for a given n, the drift function $v^{(n)}(\cdot)$ is defined on the set $\mathcal{X}^{(n)} = \{x \mid x_1 \geq -\psi_{11}\sqrt{n} - \psi_{12}\sqrt{n},\ x_2 \geq -\psi_{22}\sqrt{n}\}$, which monotonically increases and converges to $\mathbb{R}^2$. It is easy to observe that $v^{(n)}(x) \to v(x)$ uniformly on compact subsets of $\mathbb{R}^2$, where $v_2(x) = -\mu_{22}x_2$ and $v_1(x) = -\mu_{12}[x_1 \wedge (b - x_2)]$. In fact, even stronger property holds: on any compact subset of $\mathbb{R}^2$, $v^{(n)}(x) = v(x)$ for all large n.

COROLLARY 4. *The following convergence holds*

$$\hat{X}^{(n)}(\infty) \Rightarrow \hat{X}(\infty), \tag{19}$$

where $\hat{X}(\cdot)$ is a diffusion process which is a strong solution of SDE

$$d(\hat{X}) = v(\hat{X})dt + (\sigma_1 dW_1, \sigma_2 dW_2), \tag{20}$$

where W_1, W_2 are independent standard Brownian motions and the diffusion coefficients are $\sigma_1 = [\lambda_1 + \psi_{11}\mu_{11} + \psi_{12}\mu_{12}]^{1/2}$, $\sigma_2 = [\lambda_2 + \psi_{22}\mu_{22}]^{1/2}$.

The proof is fairly straightforward, we just give an outline. First, the following convergence on a finite interval holds (see e.g. [13]). Namely, consider a sequence of processes $\hat{X}^{(n)}(\cdot)$ with fixed initial states $\hat{X}^{(n)}(0) \to x \in \mathbb{R}^2$. Then, for any fixed $T_0 > 0$

$$(\hat{X}^{(n)}(t),\ t \in [0, T_0]) \Rightarrow (\hat{X}(t),\ t \in [0, T_0]), \tag{21}$$

where $\hat{X}(\cdot)$ is a strong solution of (20) with initial state $\hat{X}(0) = x$. Then, (19) can be established, together with the existence and uniqueness of a stationary distribution of $\hat{X}(\cdot)$, as follows. We consider the sequence of stationary versions of the processes $\hat{X}^{(n)}(\cdot)$ on a fixed finite time interval $[0, T_0]$, and let $n \to \infty$. Given tightness of stationary distributions of pre-limit processes, we can choose a subsequence along which $\hat{X}^{(n)}(0) \Rightarrow \tilde{X}(0)$ for some random vector $\tilde{X}(0)$; then we also have $\hat{X}^{(n)}(T_0) \Rightarrow \tilde{X}(0)$. We then use (21) to

show that the distribution of $\tilde{X}(0)$ must be a stationary distribution of $\hat{X}(\cdot)$. The uniqueness of the latter stationary distribution is easy to establish, for example, using a coupling argument.

4. Lyapunov function construction and a tightness criterion. The model in this section is quite general (including the N-system as a special case). For this model we define DFLs, construct a functional of DFL, and give sufficient conditions under which this functional can serve as a Lyapunov function to prove tightness of stationary distributions. The section is self-contained, because its main construction and result may be of independent interest. However, it may help the reader to keep the N-system described in Section 3 in mind as an example, to make the material more concrete.

4.1. *Setting and assumptions.* Let $I \geq 1$ be a fixed positive integer. For each positive integer $n \geq n_0$ (where number n_0 is fixed), we consider a Markov chain $\hat{X}^{(n)}(t),\ t \geq 0$, with a countable state space $\mathcal{S}^{(n)}$ which has the form

$$\mathcal{S}^{(n)} = \{L_n x \mid x \in \mathbb{Z}^I\} \cap \mathcal{X}^{(n)},$$

where $\mathcal{X}^{(n)}$ is a convex closed subset of $\mathbb{R}^I$, containing 0, and $L_n x = x/\sqrt{n} + s^{(n)}$ with some fixed $s^{(n)} \in \mathbb{R}^I$. Assume that for each n this Markov chain is irreducible, positive recurrent, and is such that the total transition rate out of any state is upper bounded by $R_1 n$ and any single transition has the jump size of at most $R_2/\sqrt{n}$, where R_1, R_2 are positive constants independent of n. Suppose that, defined on $\mathcal{X}^{(n)}$ is a drift function (vector field) $v^{(n)}(x)$, which is Lipschitz continuous uniformly in n. Assume that *at each point* $x \in \mathcal{S}^{(n)}$, $v^{(n)}(x)$ *gives exactly the average drift of the process.* Namely,

$$v^{(n)}(x) = \sum_{x'} (x' - x)\nu^{(n)}(x, x'), \tag{22}$$

where $\nu^{(n)}(x, x')$ is the Markov process transition rate from state x to state x'; given the upper bound on a single jump size, note that there is only a finite number of "neighbor" states x' for which $\nu^{(n)}(x, x') > 0$.

Assume that for any $x \in \mathcal{X}^{(n)}$, there is a unique solution $y^{(n)}(t),\ t \geq 0$, to the ODE

$$(d/dt)y^{(n)}(t) = v^{(n)}(y^{(n)}(t)), \quad y^{(n)}(0) = x,$$

and the solution stays within $\mathcal{X}^{(n)}$. This solution is called the *drift-based fluid limit* (DFL), starting from x.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

Suppose a continuous non-negative function $g(x),\ x \in \mathbb{R}^I$, is fixed. For $x \in \mathcal{X}^{(n)}$ define

$$G^{(n)}(x) = \int_0^\infty g(y^{(n)}(t))dt, \quad y^{(n)}(0) = x, \tag{23}$$

where $y^{(n)}(\cdot)$ is the DFL starting from x.

4.2. *A tightness criterion.* Denote by $\nabla_z G^{(n)}(x)$ the directional derivative of $G^{(n)}$ at $x \in \mathcal{X}^{(n)}$ in the direction of vector $z \in \mathbb{R}^I$:

$$\nabla_z G^{(n)}(x) \doteq \lim_{\delta \downarrow 0} \frac{1}{\delta}[G^{(n)}(x + z\delta) - G^{(n)}(x)],$$

when the limit exists. (To be precise, if x in on the boundary of $\mathcal{X}^{(n)}$, it is also required that the direction z from x points into $\mathcal{X}^{(n)}$.) Then, $\nabla_{z_*}[\nabla_z G^{(n)}](x)$ is the second derivative, first in the direction z and then z_*.

THEOREM 5. *Suppose that for any $C_1 > 0$, there exists a function $g(x),\ x \in \mathbb{R}^I$, and a constant $C_2 > 0$, such that the following conditions (i) and (ii) hold.*

(i) Function $g(x)$ is Lipschitz continuous, non-negative and such that $g(x) \to \infty$ as $x \to \infty$.

(ii) For any n, the function $G^{(n)}(x),\ x \in \mathcal{X}^{(n)}$, is finite for all x, and it has continuous gradient $\nabla G^{(n)}(x)$; for any n, any x and any fixed unit-length vectors $z, z_* \in \mathbb{R}^I$,

$$\limsup_{\delta \downarrow 0} \frac{1}{\delta} \left| \nabla_z G^{(n)}(x + z_*\delta) - \nabla_z G^{(n)}(x) \right| \le C_1 g(x) + C_2; \tag{24}$$

$$G^{(n)}(x) \to \infty, \quad x \to \infty, \qquad \textit{uniformly in } n. \tag{25}$$

Then,

$$\sup_n \mathbb{E} g(\hat{X}^{(n)}(\infty)) < \infty.$$

The second derivative condition (24) is the key one. It implies that $|\nabla_{z_*}[\nabla_z G^{(n)}](x)| \le C_1 g(x) + C_2$ if this second derivative exists. An equivalent form of (24) is as follows: for any compact set $D \subseteq \mathcal{X}^{(n)}$ and any unit-length vector $z \in \mathbb{R}^I$, the first derivative $\nabla_z G^{(n)}$ within D is Lipschitz continuous with constant

$$C_1 \max_D g(x) + C_2.$$

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

PROOF OF THEOREM 5. Let us fix a constant $C_1 > 0$, and then the corresponding function g and constant C_2, so that (i) and (ii) hold. (We will specify the choice of C_1 below.)

By definition of $G^{(n)}$ and its assumed continuous differentiability,

$$\nabla G^{(n)}(x) \cdot v^{(n)}(x) = \nabla_{v^{(n)}(x)} G^{(n)}(x) = -g(x). \tag{26}$$

Let $A^{(n)}$ denote the (infinitesimal) generator of the Markov process $X^{(n)}$. (See e.g. [6], Sections 4.1 and 1.1, for the definitions of the operator semigroup and corresponding generator of a Markov process. In our case, the semigroup is defined on the Banach space of bounded functions $h(x), x \in \mathcal{S}^{(n)}$, with norm $\sup_x |h(x)|$.) For any fixed $k > 0$, the function $G^{(n),k} \doteq G^{(n)} \wedge k$ is such that it has constant value k for all states $x \in \mathcal{S}^{(n)}$ except a finite subset $\mathcal{S}^{(n),k}$ (where the value is less than k). Therefore, there is only a finite number of possible state transitions that may change the value of $G^{(n),k}(\hat{X}^{(n)}(t))$, namely the transitions to or from the states in $\mathcal{S}^{(n),k}$; the rates of such transitions are obviously uniformly bounded. Using this property, it is easy to verify directly that function $G^{(n),k}$ is within the domain of $A^{(n)}$, that is

$$\begin{aligned}\lim_{t\downarrow 0}(1/t)E_x[G^{(n),k}(\hat{X}^{(n)}(t)) - G^{(n),k}(x)] &= A^{(n)}G^{(n),k}(x)\\ &= \sum_{x'}[G^{(n),k}(x') - G^{(n),k}(x)]\nu^{(n)}(x' - x),\end{aligned}$$

where E_x denotes the expectation conditioned on $\hat{X}^{(n)}(0) = x$, and the limit is uniform in x. This in turn implies

$$\mathbb{E}A^{(n)}G^{(n),k}(\hat{X}^{(n)}(\infty)) = 0. \tag{27}$$

(See also [9], page 31, for this property and argument in a very similar setting.) For any $x \in \mathcal{S}^{(n),k}$ we have

$$A^{(n)}G^{(n),k}(x) \le \nabla G^{(n)}(x) \cdot v^{(n)}(x) + r^{(n)}(x)(1/2)h^{(n)}(x)(R_2/\sqrt{n})^2,$$

where $R_2/\sqrt{n}$ is the maximum possible size of one jump of the process, $r^{(n)}(x) \le R_1 n$ is the total transition rate from state x, and the second-term coefficient $h^{(n)}(x)$ is bounded as $|h^{(n)}(x)| \le C_1[g(x) + \kappa_1] + C_2 = C_1 g(x) + C_1\kappa_1 + C_2 < \infty$. (The constant κ_1 appears here, because we need an upper bound on the second derivative in the R_2-neighborhood of point x, and we use the fact that $g(\cdot)$ is Lipschitz.) Recalling also (26), we obtain

$$A^{(n)}G^{(n),k}(x) \le -g(x) + (1/2)R_1R_2^2[C_1 g(x) + C_1\kappa_1 + C_2].$$

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

We now specify the choice of C_1: it is sufficiently small so that (for any $x \in \mathcal{S}^{(n),k}$)

$$A^{(n)}G^{(n),k}(x) \le -\epsilon g(x) + \kappa_2, \quad \text{for some } \epsilon > 0,\ \kappa_2 > 0.$$

(Function $g(\cdot)$ and constants κ_1, C_2 depend on C_1; therefore, constant κ_2 depends on the chosen C_1.). Obviously, if $x \in \mathcal{S}^{(n)} \backslash \mathcal{S}^{(n),k}$, which is equivalent to $G^{(n),k}(x) = k$ and equivalent to $G^{(n)}(x) \ge k$, then

$$A^{(n)}G^{(n),k}(x) \le 0.$$

From these bounds and (27) we obtain

$$\begin{aligned}&\mathbb{E}[-\epsilon g(\hat{X}^{(n)}(\infty)) + \kappa_2] I\{G^{(n)}(\hat{X}^{(n)}(\infty)) < k\} \\ &\quad \ge \mathbb{E}A^{(n)}G^{(n),k}(\hat{X}^{(n)}(\infty)) I\{G^{(n)}(\hat{X}^{(n)}(\infty)) < k\} \ge 0,\end{aligned}$$

or

$$\mathbb{E}g(\hat{X}^{(n)}(\infty)) I\{G^{(n)}(\hat{X}^{(n)}(\infty)) < k\} \le \kappa_2/\epsilon.$$

Letting $k \to \infty$, by monotone convergence,

$$\mathbb{E}g(\hat{X}^{(n)}(\infty)) \le \kappa_2/\epsilon,$$

where the constant in the RHS is independent of n. □

5. Proof of Theorem 2. We will prove Theorem 2 by choosing specific function $g(\cdot)$ and then verifying (in Theorem 10) the assumptions of Theorem 5 for N-system.

In this section, we study properties of DFL trajectories and their $G^{(n)}$-functionals, for a system with a fixed scaling parameter n. We will drop upper index (n) from now on. So, for example, will write simply $\mathcal{X}$ and $y(t)$ instead of $\mathcal{X}^{(n)}$ and $y^{(n)}(t)$, respectively. (However, the expressions may contain n as a variable.) From this point on in the paper, we say that C is a *universal constant* if C depends only on the system parameters λ_i, ψ_{ij}, μ_{ij}, b, but does *not* depend on scaling parameter n. (If the sequence of systems is defined as in Remark 1, then a universal constant C depends on the system parameters λ_i, ψ_{ij}, μ_{ij}, b, but *not* on n and *not* on the sequences $\psi_{ij}^{(n)}$ and $b^{(n)}$.)

5.1. *Basic DFL properties. First derivatives of DFLs and the Lyapunov function.* In this subsection we first establish some basic properties of DFLs and their directional (Gateaux) derivatives. Then we specify function $g(\cdot)$, and obtain the expressions for the first derivatives of the corresponding

function $G(\cdot)$. (All results of this subsection hold for systems far more general than N-system. In particular, they still hold for the systems under the *Leaf Activity Priority* LAP discipline in [16, 17], in the Halfin-Whitt regime; our priority discipline for the N-system is a special case of LAP.)

The DFL trajectories $y(\cdot)$ have the following structure. Recall that $v(x)$ is (uniformly in n) Lipschitz continuous on the entire $\mathcal{X}$. There is a finite number M (same for any n) of domains, indexed by $m = 0, \dots, M-1$; within each of them $v(x)$ is a given linear function. More precisely, the DFL satisfies a linear ODE

$$(d/dt)y(t) = v(y(t)) = u^m y(t) + a^m,$$

where u^m is a constant $I \times I$ matrix (same for each n), and a^m is a constant vector (depending on n). Informally speaking, a domain is determined by which service pools a fully occupied and/or which queues are non-empty.

Formally, the domains are easier to define (and think of) in terms of unscaled quantities $X_1 \geq 0$ and $X_2 \geq 0$, and unscaled pool sizes $B_1 = \psi_{11} n$ and $B_2 = \psi_{12} n + \psi_{22} n + b\sqrt{n}$. Each domain is defined by a combination of the directions of three strict inequalities:

$$X_1 < B_1 \quad \text{or} \quad X_1 > B_1, \tag{28}$$

$$X_2 < B_2 \quad \text{or} \quad X_2 > B_2, \tag{29}$$

$$X_1 + X_2 < B_1 + B_2 \quad \text{or} \quad X_1 + X_2 > B_1 + B_2. \tag{30}$$

However, we exclude two combinations, or conditions, $(X_1 < B_1,\ X_2 < B_2,\ X_1 + X_2 > B_1 + B_2)$ and $(X_1 > B_1,\ X_2 > B_2,\ X_1 + X_2 < B_1 + B_2)$, because they produce the empty set; and we replace ("merge") the conditions $(X_1 < B_1,\ X_2 > B_2,\ X_1 + X_2 > B_1 + B_2)$ and $(X_1 < B_1,\ X_2 > B_2,\ X_1 + X_2 < B_1 + B_2)$ into one condition $(X_1 < B_1,\ X_2 > B_2)$ because this condition alone determines the form of $v(x)$. So, there are $M = 5$ domains in total. The diffusion-scaling mapping L_n, defined by (15), transforms them into 5 (diffusion-scale) domains, denoted $\mathcal{X}^0, \dots, \mathcal{X}^4$. Note that the domains are defined by strict inequalities, so they do not cover the entire space $\mathcal{X}$. The domain closures are $\bar{\mathcal{X}}^1, \dots, \bar{\mathcal{X}}^5$, these do cover the entire $\mathcal{X}$. By these definitions, if a point belongs to the intersection of the closures of more than one domain, then necessarily at least one of the equalities (in terms of unscaled quantities), $X_1 = B_1$, $X_2 = B_2$, $X_1 + X_2 = B_1 + B_2$, holds.

In particular, consider the unscaled domain $(X_1 > B_1,\ X_2 < B_2,\ X_1 + X_2 < B_1 + B_2) = (X_1 > B_1,\ X_1 + X_2 < B_1 + B_2)$; it is such that there are no queues and pool 1 fully occupied. The corresponding diffusion-scaled domain is $\mathcal{X}^0 = \{x \in \mathcal{X} \mid x_1 > -\psi_{12}\sqrt{n},\ x_1 + x_2 < b\}$. In this domain $v(x) =$

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

$(-\mu_{12}x_1, -\mu_{22}x_2)$, i.e. $u^0 = diag(-\mu_{12}, -\mu_{22})$ and $a^0 = 0$, and therefore the components y_1 and y_2 evolve independently. Moreover, there exists a universal constant $\alpha > 0$, such that if $y(t)$ starts from a point $y(0) \in \mathcal{X}^{0,\alpha} \doteq \{\|x\| \le \alpha\} \subset \mathcal{X}^0$, then $y(t)$ never leaves domain $\mathcal{X}_0$, which in turn means that the trajectory is simply

$$y_i(t) = y_i(0)e^{-\mu_{i2}t}, \quad i = 1, 2.$$

From now such constant α and the corresponding sub-domain $\mathcal{X}^{0,\alpha}$ will be fixed.

Consider one more unscaled domain ($X_1 > B_1,\ X_2 < B_2,\ X_1 + X_2 > B_1 + B_2$). Here, pool 1 is fully occupied by type 1, pool 2 is fully occupied by X_2 type 2 customers and $B_2 - X_2$ type 1 customers, and $X_1 - B_1 - (B_2 - X_2) = X_1 + X_2 - B_1 - B_2 > 0$ type 1 customers waiting in the queue. On the diffusion scale, the domain (let us label it $m = 1$) is: $\mathcal{X}^1 = \{x \in \mathcal{X} \mid x_1 > -\psi_{12}\sqrt{n},\ x_2 < \psi_{12}\sqrt{n} + b,\ x_1 + x_2 > b\}$, and we have

$$v(x) = (\ (-b + x_2)\mu_{12},\ \ -\mu_{22}x_2),$$

with the corresponding u^1 and a^1. For the remaining 3 domains the $v(x)$ is determined similarly.

The equations for a DFL $y(\cdot)$ can be summarized as follows. The trajectory of y_2 is not affected by y_1 and satisfies ODE

$$(d/dt)y_2 = -\mu_{22}[y_2 \wedge (\psi_{12}\sqrt{n} + b)]. \tag{31}$$

If $y_1 \le -\psi_{12}\sqrt{n}$ (which corresponds to unscaled condition $X_1 \le B_1$),

$$(d/dt)y_1 = -\mu_{11}(y_1 + \psi_{12}\sqrt{n}) + \mu_{12}\psi_{12}\sqrt{n} \ge \mu_{12}\psi_{12}\sqrt{n}. \tag{32}$$

If $y_1 \ge -\psi_{12}\sqrt{n}$ ($X_1 \ge B_1$) and $y_2 \ge \psi_{12}\sqrt{n} + b$ ($X_2 \ge B_2$),

$$(d/dt)y_1 = \mu_{12}\psi_{12}\sqrt{n}. \tag{33}$$

If $y_1 \ge -\psi_{12}\sqrt{n}$ ($X_1 \ge B_1$), $y_2 \le \psi_{12}\sqrt{n} + b$ ($X_2 \le B_2$), and $y_1 + y_2 \le b$ ($X_1 + X_2 \le B_1 + B_2$), that is in domain $\bar{\mathcal{X}}^0$,

$$(d/dt)y_1 = -y_1\mu_{12}. \tag{34}$$

If $y_1 \ge -\psi_{12}\sqrt{n}$ ($X_1 \ge B_1$), $y_2 \le \psi_{12}\sqrt{n} + b$ ($X_2 \le B_2$), and $y_1 + y_2 \ge b$ ($X_1 + X_2 \ge B_1 + B_2$), that is in domain $\bar{\mathcal{X}}^1$,

$$(d/dt)y_1 = (-b + y_2)\mu_{12}. \tag{35}$$

For a given fluid trajectory, let us call time point $t \ge 0$ a *switching point* if $y(t)$ belongs to the intersection of two or more closed domains $\bar{\mathcal{X}}^m$. (i.e. it is on a boundary separating different domains).

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

LEMMA 6. *For some universal constants $T > 0$, $C' > 0$ and (integer) $K' > 0$, DFL trajectories $y(\cdot)$ satisfy the following conditions. [For a DFL $y(\cdot)$, $x = y(0) \in \mathcal{X}$ denotes its initial state.]*

(i) Let $\tau \geq 0$ be the first time a DFL reaches set $\mathcal{X}^{0,\alpha}$. Then, $\tau \leq T\|x\|$. (This, in particular, means that $y(t) \to 0$, $t \to \infty$, and, moreover, the convergence is exponentially fast.) In addition, $\|y(\cdot)\| \leq C'\|x\|$.

(ii) DFL $y(\cdot)$ depends on its initial state x continuously, in the sense of $\|y(\cdot)\|$-norm.

(iii) DFL $y(\cdot)$ has at most K' switching points, $t_1 < t_2 < \ldots < t_K$, $0 \leq K \leq K'$, and $t_K < \|x\|T$. Moreover, the set of switching points is upper semicontinuous in x; namely, as $x \to x^*$, the limiting points of the set of switching points are within the set of switching points for initial state x^*.

(iv) For any interval $[C_3, C_4]$, not containing 0, there exists a constant $T_3 > 0$ (independent of n), such that the total time the condition $y_i(t) \in [C_3, C_4]$ holds for at least one i, is upper bounded by T_3.

PROOF OF LEMMA 6. Given equation (31), condition $y_2(t) = \psi_{12}\sqrt{n} + b$ ($X_2 = B_2$) can hold at most at one point $t_2 \geq 0$, which will be a switching point. Similarly, by (32), there is at most one point $t_1 \geq 0$, at which condition $y_1(t) = -\psi_{12}\sqrt{n}$ (corresponding to $X_1 = B_1$) can hold, and if so, it will be a switching point.

Denote $t' = t_1 \vee t_2$. It is easy to see that for some universal constant $\kappa > 0$,

$$t' \leq \kappa\|x\|, \quad \|y(t')\| \leq \kappa\|x\|. \tag{36}$$

Indeed, $|y_2(t)|$ is non-increasing in $[0, \infty)$, and $t_2 \leq |x_2|/[(\psi_{12}\sqrt{n} + b)\mu_{22}] \leq |x_2|/[\psi_{12}\mu_{22}\sqrt{n}]$. In the interval $[0, t_1]$, $y_1(t)$ is negative non-decreasing, and then $|y_1(t)|$ is non-increasing; and $t_1 \leq |x_1|/[\psi_{12}\mu_{12}\sqrt{n}]$. If $t_2 > t_1$, then in the interval $[t_1, t_2]$, $(d/dt)y_1(t) = \psi_{12}\mu_{12}\sqrt{n}$, and therefore $|y_1(t_2) - y_1(t_1)| \leq \psi_{12}\mu_{12}\sqrt{n}t_2$; given the bound on t_2, we see that $|y_1(t_2) - y_1(t_1)|$ is upper bounded by $|x_2|$ times a universal constant. These observations imply (36).

For all $t > t'$, conditions $y_2(t) < \psi_{12}\sqrt{n} + b$ ($X_2 < B_2$) and $y_1(t) > -\psi_{12}\sqrt{n}$ ($X_1 > B_1$) hold. Therefore, $y(t)$ can be only in one of the two domains $\bar{\mathcal{X}}^0$ or $\bar{\mathcal{X}}^1$, depending on whether $y_1 + y_2 \leq b$ (no queues) or $y_1 + y_2 \geq b$ (queue size $y_1 + y_2 - b$ of type 1). It is easy to see from equations $(d/dt)y_2 = -\mu_{22}y_2$, (34), (35), that if $y(t)$ is in $\bar{\mathcal{X}}^1$, then the trajectory eventually leaves $\bar{\mathcal{X}}^1$ and can never return. This implies that at most two transitions between $\mathcal{X}^0$ and $\mathcal{X}^1$ can occur after t'. Specifically, either the trajectory stays in $\mathcal{X}^0$, or it is in $\mathcal{X}^1$ and then $\mathcal{X}^0$, or it is in $\mathcal{X}^0$ then $\mathcal{X}^1$

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

then $\mathcal{X}^0$. The boundary cases are also possible; for example, the trajectory may stay in the open domain $\mathcal{X}^0$ at all times, except at exactly one point $t \geq t'$ it "touches" the boundary, i.e. $y_1 + y_2 = b$. To summarize, after t' there are at most two switching points.

Denote by t'' the first time $t \geq t'$ when $\|y_2(t)\| \leq \alpha/4$. We have $t'' - t' = 0 \vee (1/\mu_{22}) \log[\|y_2(t')\|/(\alpha/4)] \leq \kappa_1 \|x\| + \kappa_2$, for some universal κ_1 and κ_2. (κ_2 depends on α, which in turn is universal.) In the interval $[t', t'']$ the value of $|y_1|$ cannot increase by more than $\kappa_3 |y_2(t')| \leq \kappa_4 \|x\|$, for universal $\kappa_3, \kappa_4 > 0$. (If $y_1 \leq 0$, then $(d/dt)y_1 \geq 0$. If $y_1 \leq 0$, then $(d/dt)y_1 \leq \mu_{12}|y_2|$, and recall that $(d/dt)y_2 = -\mu_{22} y_2$.) Therefore, $|y_1(t'')| \leq \kappa_5 \|x\|$, for universal $\kappa_5 > 0$. Starting t'', if type 1 has non-zero queue, $(d/dt)|y_1| = (d/dt)y_1 \leq -\kappa_6 < 0$, for universal $\kappa_6 > 0$; and if type 1 does not have queue, then $(d/dt)|y_1| = -\mu_{12}|y_1|$. Consider the first time $t''' \geq t''$ when $|y_1| \leq \alpha/4$. We conclude that $t''' \leq T\|x\| + \kappa_7$ and $\sup_{[0,t''']} \|y(t)\| \leq C'\|x\|$ for some universal positive constants T, C', κ_7. Obviously, $t''' \geq \tau$, so that $\tau \leq T\|x\| + \kappa_7$. However, if $\|x\| \leq \alpha$, i.e. $y(0) = x$ is already in $\mathcal{X}^{0,\alpha}$, then obviously $\tau = 0$. Therefore, in the bound $\tau \leq T\|x\| + \kappa_7$, we can drop κ_7 by rechoosing T, if necessary.

For future reference, we also make the following observation. *Suppose, $\mu_{12} = \mu_{22}$. Then, there can be at most one switching point after time t', let us call it $t_3 \geq t'$, and it is such that $y(t) \in \mathcal{X}^0$ for all $t > t_3$.* Indeed, in this case, in the domain $\bar{\mathcal{X}}^0$, we have simply $(d/dt)[y_1 + y_2] = -\mu_{22}[y_1 + y_2]$.

Let us prove properties (i)–(iv). In fact, (i) has been proved already. For a given x, let us choose τ' such that $\tau < \tau'$ for all initial states sufficiently close to x. (On a finite interval $[0, \tau']$, $y(\cdot)$ depends on the initial state continuously, because it is a solution to an ODE with Lipschitz continuous RHS.) But, for $t \geq \tau'$, the DFL with any initial state close to x is such that $y(t) \in \mathcal{X}^{0,\alpha}$; this implies uniform convergence across all $t \geq 0$, which proves (ii). The part of property (iii), stating that there is at most K' switching points, all of which are smaller than $\tau \leq T\|x\|$, has already been proved, in fact we specified that $K' \leq 4$. Then, the upper semicontinuity of the set of switching points follows from continuity of trajectories w.r.t. initial state; this proves (iii). Consider a fixed interval $[C_3, C_4]$, not containing 0. It is clear from (31) that $y_2(t)$ can spend only a finite time within $[C_3, C_4]$. Now, $y_1(t)$ can be in $[C_3, C_4]$ only after time t_1, and then in every domain the trajectory visits $y_1(t)$ satisfies one of the equations (33)–(35). If we examine each of these equations (and recall that (35) holds within domain $\bar{\mathcal{X}}^1$, where $(d/dt)y_2 = -\mu_{22} y_2$), we see that *even if the equation were to hold up to infinite time*, $y_1(t)$ can spend only a finite time within $[C_3, C_4]$. And there is only a finite, uniformly bounded number of domains that a trajectory can visit. This proves (iv). □

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

Next, let us consider the first-order dependence of DFL on the initial state. Let $y(t;x)$ denote $y(t)$ with initial state $y(0)=x\in\mathcal{X}$. For any $x\in\mathcal{X}$ and any direction $z\in\mathbb{R}^I$ (which does not point outside $\mathcal{X}$), we use the following notation for the directional (Gateaux) derivative of $y(t;x)$ at x in the direction z:

$$\nabla_z y(\cdot;x)\doteq\lim_{\delta\downarrow 0}\frac{1}{\delta}[y(\cdot;x+z\delta)-y(\cdot;x)].$$

THEOREM 7.

(i) For any fixed $x\in\mathcal{X}$ and a fixed vector z, the directional derivative

$$\xi(\cdot)=\xi(\cdot;x,z)=\nabla_z y(\cdot;x)$$

exists. It has the following structure. Let $0<t_1<t_2<\ldots<t_K$ be the switching points of $y(\cdot;x)$. Then, $\xi(0)=z$, and in each interval $[0,t_1],[t_1,t_2],\ldots,[t_K,\infty)$, ξ satisfies linear homogeneous ODE

$$(d/dt)\xi=u^m\xi,$$

where matrix u^m is the matrix u for the domain $\bar{\mathcal{X}}^m$ containing $y(t;x)$. Solutions $q(t)$, $t\geq 0$, to the equation $(d/dt)q=u^m q$, for any m, are such that

$$\|q(\cdot)\|\leq C_5\|q(0)\| \tag{37}$$

for a universal constant $C_5>0$.

(ii) The derivative $\xi(\cdot;x,z)$ depends on (x,z) continuously.

(iii) There exists a universal constant $C_6>0$, such that

$$\|\xi(\cdot;x,z)\|\leq C_6\|\xi(0;x,z)\|=C_6\|z\|.$$

PROOF. The proof of (i) relies on the following observations.

(a) In any time interval, where both $y(t;x+z\delta)$ and $y(t;x)$ are within same domain $\bar{\mathcal{X}}^m$, they are governed by the same ODE $(d/dt)y=v^m(y)$, and therefore their difference $\Delta y(t)=y(t;x+z\delta)-y(t;x)$, is governed by the linear homogeneous ODE $(d/dt)\Delta y=u^m\Delta y$. Moreover, it is easy to check that within any domain $\bar{\mathcal{X}}^m$ the corresponding matrix u^m is such that $\|\Delta y(t)\|$ can increase at most by some universal factor C_8. Indeed, consider Δy_2 first, and then Δy_1. The equation for Δy_2 is either

$$(d/dt)\Delta y_2=-\mu_{22}\Delta y_2 \tag{38}$$

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

or

$$(d/dt)\Delta y_2 = 0; \tag{39}$$

in either case $|\Delta y_2|$ cannot increase. The equation for Δy_1 is

$$(d/dt)\Delta y_1 = u_{11}^m \Delta y_1 + u_{12}^m \Delta y_2,$$

where $u_{11}^m = 0$ or $u_{11}^m = -\mu_{11}$ or $u_{11}^m = -\mu_{12}$; we also note that if Δy_2 satisfies (39) then necessarily $u_{12}^m = 0$. We see that in any case, in any time interval, $|\Delta y_2(t)|$ is upper bounded by the initial $\|\Delta y_2\|$ times a universal constant. This observation, in particular, proves (37).

(b) The total length of "switching intervals", where $y(t; x+z\delta)$ and $y(t;x)$ belong to different domains vanishes as $\delta \to 0$ (by upper semicontinuity of the set of switching points), and therefore the total change of $\Delta y(t)$ within those intervals is "small". More precisely, let t be fixed and $[\theta_1, \theta_2]$ be a switching interval such that $\theta_1, \theta_2 \to t$. Then, $\|\Delta y(\theta_2) - \Delta y(\theta_1)\|/\|\Delta y(\theta_1)\| \to 0$, because $v(x)$ is Lipschitz.

Combining observations (a) and (b), and further observing that the number of intervals where both $y(t; x + z\delta)$ and $y(t;x)$ are within same domain $\bar{\mathcal{X}}^m$ (i.e. outside the switching intervals) is upper bounded, we take the $\delta \downarrow 0$ limit to obtain (i).

(ii) This follows from the upper semicontinuity of the set of switching points on x.

(iii) By (37), in any domain $\|\xi(t)\|$ can increase at most by some factor C_5. There is only a finite number of domains that $y(t)$ visits. This proves (iii). □

We now introduce a specific function g, which we will use in the definition (23) of the Lyapunov function.

DEFINITION 8. *Let parameter $C > 0$ be fixed. Let a function $f(\eta)$ of real η be fixed, which satisfies the following conditions. It is a non-negative, even, convex, twice continuously differentiable, $f(\eta) = 0$ for $\eta \in [-C, C]$, $f'(\eta) = -1$ for $\eta \le -C - 1$, $f'(\eta) = 1$ for $\eta \ge C + 1$. (Such a function can be defined explicitly. Since C is a parameter, essentially, we fix the shape of function $f(C + \zeta)$, $\zeta \ge 0$.) Note that both f' and f'' are uniformly bounded, and $f'' = 0$ outside of the intervals $[-C-1, -C]$ and $[C, C+1]$. Then, let*

$$g(x) = \sum_i f(x_i).$$

Obviously, $|f(\eta) - |\eta|\ |$ is uniformly bounded by a constant, and then so is $|g(x) - \|x\|\ |$.

Then, by (23) we have $G(x) = \sum_i G_i(x)$, where

$$G_i(x) = \int_0^\infty f(y_i(t))dt, \quad y(0) = x. \tag{40}$$

Clearly, $G(x)$ is finite for any x, because (by Lemma 6(i)) $y(t) \to 0$ and therefore (by definition of $f(\cdot)$) $f(y_i(t)) = 0$ for all large t.

THEOREM 9. *For each i the following holds. For any $x \in \mathcal{X}$ and any direction vector z,*

$$\nabla_z G_i(x) = \int_0^\infty f'(y_i(t;x))\xi_i(t;x,z)dt. \tag{41}$$

Function $\nabla_z G_i(x)$ is continuous in (x,z).

PROOF. Expression (41) follows from Theorem 7(i) and the fact that f' is continuous bounded. The continuity of $\nabla_z G_i(x)$ is obtained using Theorem 7(ii) and Lemma 6(i,ii). □

5.2. *Second derivative bounds for the Lyapunov function.*

THEOREM 10. *The assumptions of Theorem 5 hold. Specifically, for any $C_1 > 0$, there exist constants $C > 0$ and $C_2 > 0$ such that conditions (i) and (ii) in Theorem 5 hold for the function g in Definition 8 with parameter C.*

Note that for a function g satisfying Definition 8, condition (i) of Theorem 5 holds automatically. Condition (25) is also automatic given the definition of G and basic properties of DFL, namely the fact that the time for a DFL to reach a given compact set increases to infinity as $x \to \infty$. Therefore, to prove Theorem 10, it remains to prove condition (24), and it suffices to prove it separately for G_i, $i = 1, 2$ (see (40)). We will do this first for the case $\mu_{22} \neq \mu_{12}$, and then for $\mu_{22} = \mu_{12}$. (The proof of condition (24) in this section applies to the N-system, as well as its generalization described in Section 6. It does *not* apply for LAP discipline.)

For a given x and a time $\tau^* > 0$, denote by $S(\tau^*; x)$ the set of time points, consisting of τ^* and all switching points $0 \le t < \tau^*$ of the DFL $y(\cdot; x)$.

LEMMA 11. *Suppose $\mu_{22} \neq \mu_{12}$. For any $\epsilon > 0$ there exists a sufficiently large $C_7 > 0$, such that, for all sufficiently large n, the following holds for any fixed x and any unit-length vector z. Let τ_7 be the first time the DFL $y(\cdot; x)$ hits set $\{\|y\| \le C_7\}$. Then for all sufficiently small $\delta > 0$, any point in $S(\tau_7; x + z\delta)$ is within distance at most $\epsilon\delta$ from a point in $S(\tau_7; x)$.*

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

PROOF. Consider a switching point $t \in S(\tau_7; x)$ of DFL $y(\cdot) = y(\cdot; x)$. By definition of τ_7, it is such that $\|y(t)\| \geq C_7$. The switching point is on the boundary of multiple domain closures, and therefore one or more equalities

$$y_1(t) = -\psi_{12}\sqrt{n}, \quad y_2(t) = \psi_{12}\sqrt{n} + b, \quad y_1(t) + y_2(t) = b, \tag{42}$$

defining the domain boundaries, hold. If the first or second equality holds, then $|y_i'(t)|$ is large for large n. If $y_1(t) + y_2(t) = b$, then for t to be a switching point, it is necessary that $y(t) \in \bar{\mathcal{X}}^0$; then $y_1'(t) + y_2'(t) = -\mu_{12}y_1(t) - \mu_{22}y_2(t) = -(\mu_{12} - \mu_{22})y_1(t) - \mu_{22}b$; conditions $y_1(t) + y_2(t) = b$ and $\|y(t)\| = |y_1(t)| + |y_2(t)| \geq C_7$ imply that if C_7 is large then so is $|y_1(t)|$, and then $|y_1'(t) + y_2'(t)|$ is large as well. We conclude that if any of the three equalities (42) holds, then for all $n \geq n'$ we have $|y_1'(t)| \geq \kappa_1$ or $|y_2'(t)| \geq \kappa_1$ or $|y_1'(t) + y_2'(t)| \geq \kappa_1$, respectively, where the constant $\kappa_1 > 0$ can be made arbitrarily large by choosing large enough n' and C_7. This means that, first, the domains in which the trajectory $y(\cdot; x)$ is in before and after the switching point t are uniquely defined. Second, since the distance between $y(\cdot; x + z\delta)$ and $y(\cdot; x)$ does not exceed $\kappa_2\delta$ at all times, where $\kappa_2 > 0$ is a universal constant (this follows from Theorem 7), and $v(\cdot)$ is Lipschitz, any point in $S(\tau_7; x + z\delta)$ must be within $2\kappa_2\delta/\kappa_1$ of a point in $S(\tau_7; x)$. Since κ_2 is universal and κ_1 can be made arbitrarily large (by choosing C_7 large), the result follows. □

Recall that to prove Theorem 10, it remains to prove the second derivative condition (24). The proper second derivative may not exist, hence we must "settle" for the estimate (24). But, to illustrate the proof that follows, let us write down the expression for the second derivative, by formally applying ∇_{z_*} differentiation to (41) (this expression is *not* used in the proof):

$$\nabla_{z_*}\nabla_z G_i(x)$$

$$= \int_0^\infty f''(y_i(t; x))\xi_i(t; x, z_*)\xi_i(t; x, z)dt \tag{43}$$

$$+ \int_0^\infty f'(y_i(t; x))\nabla_{z_*}\xi_i(t; x, z)dt. \tag{44}$$

PROOF OF THEOREM 10, CASE $\mu_{22} \neq \mu_{12}$. We choose small $\epsilon > 0$ and then $C_7 > 0$ as in Lemma 11. Then choose parameter $C > 0$ of function g large enough so that any DFL starting from the set $\{\|y\| \leq 2C_7\}$ never hits set $\{\|y\| \geq C\}$. (We can do this by Lemma 6(i).)

For $i = 1, 2$ consider

$$\frac{1}{\delta}[f'(y_i(t;x+z_*\delta))\xi_i(t;x+z_*\delta,z) - f'(y_i(t;x))\xi_i(t;x,z)]$$

$$= \frac{1}{\delta}[f'(y_i(t;x+z_*\delta)) - f'(y_i(t;x))]\xi_i(t;x,z) \tag{45}$$

$$+ \frac{1}{\delta}f'(y_i(t;x+z_*\delta))[\xi_i(t;x+z_*\delta,z) - \xi_i(t;x,z)]. \tag{46}$$

(The integrals of the terms (45) and (46), correspond to the integrals (43) and (44), respectively, in the formal second derivative expression.)

Since $f(\cdot)$ has bounded second derivative, the term (45) converges (uniformly in t) to

$$f''(y_i(t;x))\xi_i(t;x,z_*)\xi_i(t;x,z).$$

The integral of this over $t \in [0,\infty)$ is bounded because the total time any trajectory spends in the set $\{C \le \|y_i\| \le C+1\}$ is uniformly bounded (by Lemma 6(iv).)

In the term (46), $f'(y_i(t;x+z_*\delta))$ is uniformly bounded. Let τ_7 be the first time $y(t;x)$ hits set $\{\|y\| \le C_7\}$. We claim that, uniformly in $t \in [0,\tau_9]$,

$$\limsup_{\delta\downarrow 0} \frac{1}{\delta}[\xi_i(t;x+z_*\delta,z) - \xi_i(t;x,z)] \le \epsilon\kappa, \tag{47}$$

where $\kappa > 0$ is a universal constant. Indeed, let $t_1 \in S(\tau_7;x)$ be the first (smallest) switching point of trajectory $y(\cdot;x)$. To be concrete, let us assume $t_1 > 0$. (The case $t_1 = 0$ is treated analogously.) For a given δ, we define a *switching interval* $[\theta_1^*, \theta_1^{**}]$ associated with t_1 as follows: θ_1^* is the minimum of t_1 and those switching points of $y(\cdot; x+z\delta)$ that are within distance $\epsilon\delta$ from t_1; similarly, θ_1^{**} is the maximum of t_1 and those switching points of $y(\cdot;x+z\delta)$ that are within distance $\epsilon\delta$ from t_1. Obviously, $\theta_1^{**} - \theta_1^* \le 2\epsilon\delta$. In the interval $[0,\theta_1^*]$, $\xi(t;x+z_*\delta,z) = \xi(t;x,z)$, because they are governed by the ODE with *same* matrix u^m. Within the switching interval, the ODEs for $\xi(t;x+z_*\delta,z)$ and $\xi(t;x,z)$ may have a different matrix u^m, but there is only a finite number of those matrices; therefore, in $[\theta_1^*,\theta_1^{**}]$, $\|\xi(t;x+z_*\delta,z) - \xi(t;x,z)\|$ can increase at most by $\kappa_1\|\xi(\theta_1^*;x,z)\|\epsilon\delta$, where κ_1 is a universal constant. We then consider the second switching point t_2 and the associated switching interval $[\theta_2^*,\theta_2^{**}]$. Note that between the first and second switching intervals, both $\xi(t;x+z_*\delta,z)$ and $\xi(t;x,z)$ are again governed by the ODE with same matrix u^m; therefore the difference $\xi(t;x+z_*\delta,z) - \xi(t;x,z)$ is governed by the same ODE, and therefore in the interval $[\theta_1^{**},\theta_2^*]$ the value of $\|\xi(t;x+z_*\delta,z) - \xi(t;x,z)\|$ can increase at most by a factor given by a universal constant $\kappa_2 > 0$ (by (37)). At the end of the switching interval

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

$[\theta_2^*, \theta_2^{**}]$, the first-order component of $\|\xi(t; x + z_*\delta, z) - \xi(t; x, z)\|$ is upper bounded by

$$\kappa_2\kappa_1\|\xi(\theta_1^*; x, z)\|\epsilon\delta + \kappa_1\|\xi(\theta_2^*; x, z)\|\epsilon\delta.$$

We consider the third switching point, and so on. We see that the first-order component of $\|\xi(t; x + z_*\delta, z) - \xi(t; x, z)\|$ will be upper bounded by $\kappa\epsilon\delta$, for a sufficiently large universal κ. (There will be also higher order terms δ^ℓ, $\ell \geq 2$, with uniformly bounded coefficients.) This proves claim (47).

By Lemma 6(i), $\tau_7 \leq \kappa_3\|x\|$, for a universal constant $\kappa_3 > 0$. (We can always choose $C_7 \geq \alpha$, and then $\kappa_3 = T$.) Then, $\tau_7 \leq \kappa_3(g(x) + \kappa_4)$, where κ_4 may depend on the parameter $C > 0$ of function g.

Now, for all sufficiently small δ, the integral of the term (46),

$$\begin{aligned}&\int_0^\infty \frac{1}{\delta} f'(y_i(t; x + z_*\delta))[\xi_i(t; x + z_*\delta, z) - \xi_i(t; x, z)]dt \\ &\quad = \int_0^{\tau_9} \frac{1}{\delta} f'(y_i(t; x + z_*\delta))[\xi_i(t; x + z_*\delta, z) - \xi_i(t; x, z)]dt,\end{aligned}$$

because $f'(y_i(t; x + z_*\delta)) = 0$ for $t \geq T$. The absolute value of the latter integral is upper bounded by

$$\kappa_3(g(x) + \kappa_4)\kappa\epsilon = \kappa\kappa\epsilon g(x) + \kappa_3\kappa_4\kappa\epsilon.$$

The constants κ_3 and κ are universal, while κ_4 depends on C, which depends on C_7, which depends on ϵ. It remains to choose ϵ small enough so that $\kappa_3\kappa\epsilon < C_1$. Then the value of $\kappa_3\kappa_4\kappa\epsilon$, plus the corresponding upper bound on the integral of (45), gives constant C_2. □

PROOF OF THEOREM 10, CASE $\mu_{22} = \mu_{12}$. This case is treated the same way as $\mu_{12} \neq \mu_{22}$, with the following modifications. If there is *no* switching point $t \in S(\tau_7; x)$, associated with equality $y_1(t)+y_2(t) = b$, then the proof is unchanged. Suppose there is a switching point $t \in S(\tau_7; x)$, associated with equality $y_1(t) + y_2(t) = b$. Then, in the notation of the proof of Lemma 6, we must have $t \geq t'$, and by the observation we made in that proof, t is the last switching point, and therefore it is the only switching point associated with equality $y_1(t) + y_2(t) = b$. Moreover, all the properties we established in the $\mu_{22} \neq \mu_{12}$ case proof, still apply to all switching points before t. After time t, the process stays within the domain $\mathcal{X}_0$, and therefore $(d/dt)[y_1(t) + y_2(t)] = -\mu_{22}[y_1(t)+y_2(t)]$. In particular, in a small neighborhood of time t, $(d/dt)[y_1(t) + y_2(t)] \leq -(b/2)\mu_{22} < 0$. These facts imply that the switching interval, corresponding to switching time t, is such that its end points are within $\kappa_5\delta$ from t, for some universal constant $\kappa_5 > 0$. This means that the

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

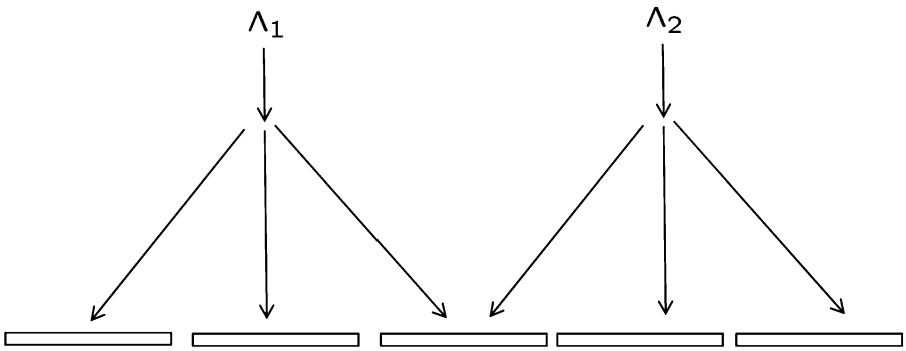


FIG 3. *A more general system.*

contribution of this last switching interval, as well as of the remaining time interval up to the time τ_7, into the integral of (46), is upper bounded by a universal constant $\kappa_6 > 0$. □

6. Generalization of the *N*-system. Theorem 2, along with its proof, easily extend to the generalization of N-system, shown in Figure 3, in the Halfin-Whitt regime. The system has two customer types and arbitrary number of server pools. There is exactly one server pool that is flexible, i.e. can serve both types. (On Figure 3, it is the pool in the middle.) Each of the remaining pools is dedicated to service of either type 1 or 2. (The two pools on the left in the figure are dedicated to type 1, while the two pools on the right – to type 2.) Each customer type has absolute preference for its dedicated server pools, in some fixed priority order, over the flexible pool. In the flexible pool, the absolute preemptive priority is given to one of the types.

The key features that the generalized system shares with the N-system are that there are two customer types and only one flexible server pool, which can be shared by the customers of different types. These features are exploited in Section 5.2, where we estimated second derivatives of the Lyapunov function. (We note again that all results in Section 5.1, which concern with first derivatives, hold for far more general systems, e.g. those under LAP discipline [16, 17].) The behavior of the DFLs for the generalized system is more complicated, simply because the number of state space domains can be very large. However, as in the N-system, after a finite time all dedicated server pools stay fully occupied, which means that the DFL dynamics depends only on "what happens" in the flexible pool. Consequently, our analysis goes through with very minor adjustments.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

7. Discussion. In this paper we address the problem of tightness of stationary distributions, and the limit interchange, for flexible multi-pool service systems in the Halfin-Whitt regime. The behavior of such systems can be very complicated, which makes the problem challenging. This is, in particular, due to the difficulty of constructing Lyapunov functions. Our approach uses a (family of) Lyapunov function(s), defined as an integral functional of the drift-based fluid limits (DFL) $y(\cdot)$: $G(x) = \int_0^\infty g(y(t))dt$, $y(0) = x$. The problem then reduces to studying the (first and second) derivatives of a DFL – and the corresponding integral $G(x)$ – on the initial state x. We apply this approach to show the tightness property for the N-model under a priority discipline.

Both the approach and many parts of our analysis are quite generic and might be applicable to other models as well. In this respect, note that there is a lot of flexibility in choosing the "distance" function $g(\cdot)$. It might also be possible to combine the approach with other techniques. For example, a Lyapunov function of the type we consider could be defined and applied on a subspace, if it could be shown by other means that the stationary distributions concentrate (in appropriate sense) on that subspace. Exploring these directions may be a subject of future research.

APPENDIX A: POSITIVE RECURRENCE PROOF

Let us drop superscript (n). Consider the process with fixed initial state such that $X_2(0) = 0$. Consider the sequence of time points $0 < t_1 < t_2 < \dots$ at which $X_2(t)$ changes to 0; let $t_0 = 0$. These are renewal points for $X_2(\cdot)$ viewed in isolation; $X_2(\cdot)$ is positive recurrent. The renewal interval durations $t_{\ell+1} - t_\ell$, $\ell = 0, 1, \dots$, are of course i.i.d. with some finite mean T. Let A_ℓ be the random number of type 1 arrivals into the system in the interval $(t_\ell, t_{\ell+1}]$; and S_ℓ be the random number of type 1 service completions in the interval $(t_\ell, t_{\ell+1}]$, *assuming that all servers (in both pools), not occupied by type 2 customers, serve type 1 customers.* Clearly, (A_ℓ, S_ℓ) are i.i.d. across ℓ, $\mathbb{E}A_\ell = \lambda_1 nT$, $\mathbb{E}S_\ell = \lambda_1 nT + b\mu_{12}\sqrt{n}T$, $\mathbb{E}A_\ell - \mathbb{E}S_\ell = -b\mu_{12}\sqrt{n}T < 0$. Using these facts, it is easy to see that the discrete time Markov chain $X_1(t_\ell), \ell = 0, 1, 2, \dots$ is positive recurrent; let $N < \infty$ denote the mean return time to 0 for this chain. This implies that, for the original continuous-time process $(X_1(t), X_2(t))$, the mean time to return to state $(0, 0)$ is upper bounded by NT. We omit further details.

REFERENCES

[1] Aksin, Z., Armony, M., and Mehrotra, V. (2007). The modern call-center: A multi-disciplinary perspective on operations management research. *Production and Operations Management, Special Issue on Service Operations in Honor of John Buzacott (ed. G. Shanthikumar and D. Yao)* **16**, 6, 655–688.

[2] Budhiraja, A. and Lee, C. (2009). Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Math. Oper. Res.* **34**, 45–56. MR2542988

[3] Dai, G. J., Dieker, A. B., and Gao, X. (2014). Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. *Queueing Systems* **78**, 1, 1–29. MR3238006

[4] Dieker, A. B. and Gao, X. (2013). Positive recurrence of piecewise Ornstein–Uhlenbeck processes and common quadratic Lyapunov functions. *Annals of Applied Probability* **23**, 4, 1291–1317. MR3098433

[5] Dupuis, P. and Williams, R. (1994). Lyapunov functions for semimartingale reflecting Brownian motions. *Annals of Probability* **22**, 2, 680–702. MR1288127

[6] Ethier, S. and Kurtz, T. (1986). *Markov Processes. Characterization and Convergence*. Wiley. MR0838085

[7] Gamarnik, D. and Goldberg, D. (2013). Steady-state GI/GI/n queue in the Halfin-Whitt regime. *Annals of Applied Probability* **23**, 6, 2382–2419. MR3127939

[8] Gamarnik, D. and Momcilovic, P. (2008). Steady-state analysis of a multiserver queue in the halfin-whitt regime. *Advances in Applied Probability* **40**, 548–577. MR2433709

[9] Gamarnik, D. and Stolyar, A. L. (2012). Multiclass multiserver queueing system in the halfin-whitt heavy traffic regime. asymptotics of the stationary distribution. *Queueing Systems* **71**, 25–51. MR2925789

[10] Gamarnik, D. and Zeevi, A. (2006). Validity of heavy traffic steady-state approximations in generalized jackson networks. *The Annals of Applied Probability* **16**, 56–90. MR2209336

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

[12] Gurvich, I. (2014). Validity of heavy-traffic steady-state approximations in multiclass queueing networks: The case of queue-ratio disciplines. *Mathematics of Operations Research* **39**, 1, 121–162. MR3173006

[13] Gurvich, I. and Whitt, W. (May 2009). Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of OR* **34**, 2, 363–396. MR2554064

[14] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**, 3, 567–588. MR0629195

[15] Schonlein, M. and Wirth, F. (2012). On converse Lyapunov theorems for fluid network models. *Queueing Systems* **70**, 339–367. MR2897699

[16] Stolyar, A. L. (2015). Diffusion scale tightness of invariant distributions of a large-scale flexible service system. *Advances in Applied Probability* **47**, 251–269. MR3327324

[17] Stolyar, A. L. and Yudovina, E. (2012). Tightness of invariant distributions of a large-scale flexible service system under a priority discipline. *Stochastic Systems* **2**, 2, 381–408. MR3354771

[18] Stolyar, A. L. and Yudovina, E. (2013). Systems with large flexible server pools: Instability of "natural" load balancing. *Annals of Applied Probability* **23**, 5, 2099–2138. MR3134731

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.

[19] Ye, H. Q. and Chen, H. (2001). Lyapunov method for the stability of fluid networks. *Operations Research Letters* **28**, 125–136. MR1833207

Alexander L. Stolyar
Lehigh University
Bethlehem, PA
E-mail: stolyar@lehigh.edu

Downloaded from informs.org by [216.73.217.172] on 15 November 2026, at 06:06 . For personal use only, all rights reserved.