The Annals of Applied Probability
2005, Vol. 15, No. 3, 1652–1690
DOI 10.1214/105051605000000179

INSTABILITY IN STOCHASTIC AND FLUID QUEUEING NETWORKS

BY DAVID GAMARNIK AND JOHN J. HASENBEIN[1]

IBM T. J. Watson Research Center and University of Texas at Austin

The fluid model has proven to be one of the most effective tools for the analysis of stochastic queueing networks, specifically for the analysis of stability. It is known that stability of a fluid model implies positive (Harris) recurrence (stability) of a corresponding stochastic queueing network, and weak stability implies rate stability of a corresponding stochastic network. These results have been established both for cases of specific scheduling policies and for the class of all nonidling policies.

However, only partial converse results have been established and in certain cases converse statements do not hold. In this paper we close one of the existing gaps. For the case of networks with two stations, we prove that if the fluid model is not weakly stable under the class of all nonidling policies, then a corresponding queueing network is not rate stable under the class of all nonidling policies. We establish the result by building a particular nonidling scheduling policy which makes the associated stochastic process transient. An important corollary of our result is that the condition $\rho^* \le 1$, which was proven in [*Oper. Res.* **48** (2000) 721–744] to be the exact condition for global weak stability of the fluid model, is also the exact global rate stability condition for an associated queueing network. Here ρ^* is a certain computable parameter of the network involving virtual station and push start conditions.

1. Introduction.

In a series of papers, starting in the early 1990's, researchers established a strong connection between the stability of a queueing network and the stability of the corresponding fluid model. Initiated by Rybko and Stolyar [20] and generalized by Dai [7], Stolyar [21] and Chen [4], among others, it has been demonstrated that the stability of a fluid model implies stability of a corresponding queueing network. The stability results in the aforementioned papers were established both for classes of policies, for example, the set of nonidling policies, and specific policies, for example, First-In-First-Out (FIFO) (Dai [7] discusses both types of results). The fluid model is a continuous, deterministic analog of a discrete stochastic queueing network. It is defined through a set of equations which nominally take as parameters only the mean values of the random variables associated with the queueing network.

Received October 2003; revised September 2004.

[1] Supported in part by NSF Grant DMI-01-32038.

AMS 2000 subject classifications. Primary 90B15; secondary 60K25, 60F10.

Key words and phrases. Rate stability, fluid models, large deviations.

Since the stability behavior of the fluid model is often significantly easier to analyze than that of the stochastic model, the results above have led to sweeping advances in understanding the stability of queueing networks via the fluid model. A short list of such papers includes [1, 3, 5, 11, 12, 16]. However, a major element needed for a satisfactory theory of stability via fluid models is a converse to the aforementioned stability results. Specifically, if the fluid model is *not* stable in some sense, does this imply instability of the corresponding queueing network? Unfortunately, it turns out that formulating an appropriate converse is a delicate matter. Partial converses which appear in the literature refer both to the fluid model and the fluid limit model, which is the set of weak limits of the rescaled stochastic process. Dai [8] introduces the notion of a weakly unstable fluid limit model. Roughly speaking, the fluid limit model is weakly unstable if there exists a uniform time at which all fluid limits which start at zero are strictly positive. If the fluid limit model is weakly unstable, Dai provides a concise proof showing that in the stochastic network, the queue length process diverges to infinity with probability one. This result provides a partial converse to the stability results mentioned earlier. Puhalskii and Rybko [19] use large deviations methods to prove another partial converse to the stability theorems. Their result implies that if there exists an initial fluid model state for which all fluid trajectories with “close” initial states satisfy a uniform rate of divergence condition, then the queueing process is not positive Harris recurrent. Under stronger conditions on the fluid trajectories they prove transience of the queueing process. In two different papers, Meyn focuses on networks which can be represented by countable state Markov chains. In [17], Meyn uses martingale methods to show that if all fluid limits eventually diverge at some uniform rate, then the state process associated with the queueing network is transient. Meyn [18] uses Markov chain techniques to prove another transience result. In that paper, if the fluid limits satisfy a uniform homogeneity condition and a uniform lower bound for trajectories starting from some open set, this again implies that the state process for the associated queueing network is transient. In each of the papers [17] and [18], Meyn explains how the results can be extended to networks with more general state spaces.

In all of the papers above which prove a converse result to the original stability theorems in [4, 7, 21], some uniform requirement over a set of fluid trajectories or, more precisely, a set of fluid limits (sometimes restricted to fluid limits starting from a particular type of state), is needed for the result to be applicable. Recall that the original stability results of Dai [7] and Chen [4] require that *all* fluid trajectories are stable in some sense. Hence, we use the term “partial converse” above because there is some gap between the stability and instability results. To close the gap between the stability and instability results, one might consider analyzing directly the set of fluid limits. However, this approach presents certain difficulties, since the fluid limits are defined in a nonconstructive way, as weak limits of the underlying stochastic process. Moreover, it is shown in [15] that computing fluid limits of a queueing system is an algorithmically undecidable problem for a certain class of scheduling policies. In contrast, the fluid trajectories are defined by the set of solutions of a fluid model, a series of a fairly simple and reasonably tractable differential equations. It is this latter notion of fluid model which we use herein.

In [4], it is shown that a multiclass queueing network is globally rate stable if the corresponding fluid network is globally weakly stable (see Section 2.2.2 in this paper for definitions). In this paper we present a result which is a full converse to Chen's stability result. It is a full converse in that, for some networks, in particular, two station networks, the result implies that the stochastic network is globally, rate stable if and only if the corresponding fluid network is globally weakly stable. In particular, this implies that if there is just one linearly divergent fluid trajectory, then the stochastic network is not rate stable under some nonidling policy. Combining our main result with the result of Dai and VandeVate [12], we show that a certain computable condition of the form $\rho^* \leq 1$ is a necessary and sufficient condition for rate stability in networks with two stations. This is the first tight condition for stability for such a general class of networks. Our proof uses a series of large deviations estimates to establish the result and the only restriction in the stochastic network is that the estimates are applicable to the primitive stochastic processes defining the network. For a comprehensive discussion of various stability concepts in fluid and queueing networks, and the $\rho^* \leq 1$ condition, we suggest Dai [9].

It should be noted that a strength of the transience results in [8, 17–19] is that they can be applied to networks under a class of policies or just one particular policy (like FIFO or a static buffer priority policy), whereas our result only applies to the class of nonidling policies. In other words, the advantage of the previous transience results is that they can be used to determine if a given network is stable under a particular scheduling policy. Our result can be used only to determine if there exists one scheduling policy, within the class of nonidling policies, which makes a network unstable. It should be noted though that, in general, it is more difficult to apply the previous results because of more stringent requirements on the behavior of the fluid model trajectories.

One is naturally led to ask if our result can be extended to apply to networks operating under a particular policy rather than the class of all nonidling policies. Unfortunately, a paper by Dai, Hasenbein and VandeVate [10] essentially rules out the possibility of obtaining a full converse which can also be applied to particular policies. In that paper, it is shown that the stability of a queueing network under a fixed static buffer priority policy depends on more than just the mean value of the service and interarrival times. Hence, no mean-value based fluid model can sharply determine stability for the network considered, which implies that no general stability converse can be formulated for a network operating under an arbitrary, but specific policy.

Our paper is organized as follows, in Section 2 we introduce stochastic and fluid multiclass networks and describe mathematical preliminaries. In Section 3 we present the main results of our paper and their implications. All of the proofs are presented in Section 4.

2. Preliminaries—model description and assumptions. We start by describing the model of interest—a multitype queueing network. In the following section we describe a stochastic multitype queueing network and in Section 2.2 we introduce a fluid queueing network.

2.1. *Stochastic multitype queueing networks.*

2.1.1. *Network description.* An open stochastic multitype queueing network is a network of J stations $\sigma_1, \sigma_2, \dots, \sigma_J$ each processing one or multiple types of jobs. For each type $i = 1, 2, \dots, I$, there is an external stream of jobs arriving to the network. The intervals between successive arrivals of jobs corresponding to type i are given by the i.i.d. sequence $\mathbf{X}_1^i, \mathbf{X}_2^i, \dots, \mathbf{X}_k^i, \dots$. If $\mathbb{E}[\mathbf{X}_1^i]$ exists, we define $\lambda_i \equiv 1/\mathbb{E}[\mathbf{X}_1^i]$ to be the arrival rate for type i. More detailed assumptions about the stochastic processes $\{\mathbf{X}_k^i, k = 1, 2, \dots\}$ are provided later. We denote by $\mathbf{A}_i(t)$ the cumulative arrival process which counts the number of arrivals up to time t. That is, $\mathbf{A}_i(t) = \max\{k : \sum_{r \le k} \mathbf{X}_r^i \le t\}$.

Each job of type i has to be processed on a fixed ordered sequence of stations $\sigma(i, 1), \sigma(i, 2), \dots, \sigma(i, J_i)$, where each $\sigma(i, l)$ is one of the stations $\sigma_1, \dots, \sigma_J$. We refer to $(i, 1), (i, 2), \dots, (i, J_i)$ as stages corresponding to the type i. We allow the repetition of stations, that is, $\sigma(i, j') = \sigma(i, j'')$ for $j' \neq j''$, meaning some jobs need to be processed on the same station multiple times (which is common in some manufacturing environments). In particular, J_i could be bigger than J. We slightly abuse the notation sometimes by using σ to also denote the set of classes which are served at station σ.

Each station $\sigma = \sigma_j, j \le J$, has one server and, in particular, can work on only one job at a time. Other jobs awaiting processing on σ accumulate into queues. Type i jobs in the queue corresponding to stage (i, j) will be referred to as *class* (i, j) jobs. Once a job of class (i, j) is processed, it is moved into the next queue $(i, j+1)$ at the station $\sigma(i, j+1)$, or leaves the network if $j = J_i$. The processing times for jobs of type i at stage j are random and are given by the i.i.d. sequence $\mathbf{S}_1^{i,j}, \mathbf{S}_2^{i,j}, \dots, \mathbf{S}_k^{i,j}, \dots$. If $\mathbb{E}[\mathbf{S}_1^{i,j}]$ exists, we define $\mu_{i,j} \equiv 1/\mathbb{E}[\mathbf{S}_1^{i,j}]$ to be the service rate for jobs in class (i, j). Again, more detailed assumptions regarding the stochastic processes $\{\mathbf{S}_k^{i,j}, k = 1, 2, \dots\}$ are provided later.

Let $d = \sum_i J_i$ denote the total number of classes in the network. We denote by $\mathbf{Q}(t) = (\mathbf{Q}_{i,j}(t)) \in \mathbb{Z}_+^d$ the vector of queue lengths in our queueing network at time $t \ge 0$. In order to completely specify the stochastic dynamics of $\mathbf{Q}(t)$, we need to specify the vector of initial queue lengths $q = \mathbf{Q}(0)$ and the *scheduling policy* $\mathcal{U}$ which gives gives the protocol at each station σ for resolving the contention for service, when several jobs are competing for the same station. Some common policies include the First-In-First-Out (FIFO) policy which gives priority to jobs which arrived earlier to the station, Last-In-First-Out (LIFO) defined analogously, Global-FIFO (GFIFO) which gives priority to jobs which arrived earlier into the entire network [based on time stamps of a job's arrival to class $(i,1)$] and static buffer priority policies which are based on a ranking of classes in each station and give priority to jobs with the higher ranking, and so on. All of these policies are examples of *nonidling* policies, which are defined as policies that require each station σ to work at full capacity as long as there are any jobs waiting to be processed by σ.

Throughout the paper we will be only considering head-of-the-line (HOL) type nonidling scheduling policies. Under an HOL policy, at most, one job of each class at a given station can receive service at a given time. Furthermore, under the HOL assumption, jobs are served in FIFO order within a given class. FIFO, GFIFO and static buffer priority are examples of HOL policies. Adopting the HOL assumption in this paper is really not a restriction since the main goal of this paper is to construct an unstable (in a sense to be defined) nonidling scheduling policy. Indeed, we construct an unstable policy which happens to be of HOL type. In addition to being HOL, the policy we use to prove the main result is *preemptive resume*. Under such a policy, if the processing of a job of class i is interrupted to serve a job from another class j, then the class i job is ejected from service and placed at the head of the line for processing at a later time. When class i is again chosen for service, the remaining processing time for the ejected job is the same as it was at the moment it was ejected.

For each $(q,z_1,z_2)\in\mathbb{Z}_+^d\times\Re_+^{I+d}$, we say that the state of the stochastic process at time t is (q,z_1,z_2) if at time t the vector of queue lengths $\mathbf{Q}(t)$ is q, the vector of residual interarrival times is z_1 (hence, the dimension I for this component of the state) and the vector of residual service times is z_2. For many scheduling policies, including the policy constructed in this paper, the state space $\mathbb{Z}_+^d\times\Re_+^{I+d}$ is adequate to describe the underlying stochastic process of the network.

For each class (i,j), let $\mathbf{T}_{i,j}(t)$ denote the total amount of time station $\sigma(i,j)$ spent processing class (i,j) jobs during the time interval $[0,t]$. Let $\mathbf{D}_{i,j}(t)$ denote the cumulative departure process for class (i,j) jobs, that is, $\mathbf{D}_{i,j}(t)$ is the number of class (i,j) jobs that station $\sigma_{i,j}$ processed during the time interval $[0,t]$. For each station σ, let $\mathbf{Q}_\sigma(t)=\sum_{(i,j)\in\sigma}\mathbf{Q}_{i,j}(t)$ and let $\mathbf{T}_\sigma(t)=\sum_{(i,j)\in\sigma}\mathbf{T}_{i,j}(t)$. The following relations follow immediately from the definitions. For all $1\le i\le I$, $2\le j\le J_i$ and $t\ge 0$,

$$\mathbf{Q}_{i,1}(t)=\mathbf{Q}_{i,1}(0)+\mathbf{A}_i(t)-\mathbf{D}_{i,1}(t), \tag{1}$$

$$\mathbf{Q}_{i,j}(t)=\mathbf{Q}_{i,j}(0)+\mathbf{D}_{i,j-1}(t)-\mathbf{D}_{i,j}(t), \tag{2}$$

$$\mathbf{D}_{i,j}(t)=\max\left\{k:\sum_{r\le k}\mathbf{S}_r^{i,j}\le\mathbf{T}_{i,j}(t)\right\}. \tag{3}$$

Also for every $0\le t_1\le t_2$ and every station σ,

$$\sum_{(i,j)\in\sigma}\bigl(\mathbf{T}_{i,j}(t_2)-\mathbf{T}_{i,j}(t_1)\bigr)\le t_2-t_1. \tag{4}$$

Finally, if the scheduling policy $\mathcal{U}$ is nonidling, then for every $0 \le t_1 \le t_2$ and every station σ, if $\mathbf{Q}_\sigma(t) > 0$ for all $t_1 \le t \le t_2$, then $\mathbf{T}_\sigma(t_2) - \mathbf{T}_\sigma(t_1) = t_2 - t_1$. In other words, if the total queue in station σ was always positive during the time interval $[t_1, t_2]$, then the station was always working on jobs full time during this interval.

Let

$$\lambda_{\max} = \max_i\{\lambda_i, \lambda_i^{-1}\}, \tag{5}$$

$$\mu_{\max} = \max_{i,j}\{\mu_{i,j}, \mu_{i,j}^{-1}\}, \tag{6}$$

$$J_{\max} = \max_i\{J_i\}. \tag{7}$$

For technical purposes, we introduce C—a very large constant which exceeds all the parameters of the network. Specifically,

$$C > 13(\lambda_{\max} + \mu_{\max})^2 I J_{\max}^3. \tag{8}$$

For any station σ, let $|\sigma|$ denote the number of classes in the set σ. For any vector $q \in \Re^d$, we let $\|q\| = \sum_{1\le i\le d} |q_i|$ denote the L_1 norm. For any nondecreasing nonnegative function $f(t)$ and any $t_1 \le t_2$, we let $f(t_1, t_2)$ denote $f(t_2) - f(t_1)$.

2.1.2. *Stochastic assumptions.* Below, we introduce some basic assumptions on the sequences of random variables which represent the primitive data in our stochastic networks, and an assumption on the behavior of the network process itself.

DEFINITION 1. Consider a sequence of i.i.d. nonnegative random variables $\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n, \dots$ with $\mathbb{E}[\mathbf{Z}_1] \equiv \alpha < \infty$. Such a sequence satisfies large deviations (LD) bounds if for every $\varepsilon > 0$, there exist constants $L = L(\varepsilon)$, $V = V(\varepsilon) > 0$ such that, for any $z > 0$,

$$\mathbb{P}\left(\left|\sum_{1\le i\le n} \mathbf{Z}_i - z - \alpha n\right| \ge \varepsilon n \,\middle|\, \mathbf{Z}_1 \ge z\right) \le V e^{-Ln}, \tag{9}$$

for all $n \ge 1$, and the counting process $\mathbf{N}(t) \equiv \max\{n : \mathbf{Z}_1 + \cdots + \mathbf{Z}_n \le t\}$ satisfies

$$\mathbb{P}\left(\left|\mathbf{N}(t+z) - \frac{t}{\alpha}\right| \ge \varepsilon t \,\middle|\, \mathbf{Z}_1 \ge z\right) \le V e^{-Lt}, \tag{10}$$

for all $t \ge 0$.

It is important that the constants L, V in the definition above do not depend on $z > 0$. This uniformity will become useful when we analyze arrival and service processes with the presence of some residual interarrival and service times. For simplicity, we assume common constants $L = L(\varepsilon), V = V(\varepsilon)$ instead of individual constants corresponding to indices i, j. Our main stochastic assumptions are as follows:

ASSUMPTION A. The sequences $\{\mathbf{X}_n^i, n = 1, 2, \dots\}$ and $\{\mathbf{S}_n^{i,j}, n = 1, 2, \dots\}$ are i.i.d. for every $1 \le i \le I$ and $1 \le j \le J_i$.

ASSUMPTION B. For each i, j, the large deviation bound (9) holds for the sequences $\{\mathbf{X}_n^i, n = 1, 2, \dots\}$ and $\{\mathbf{S}_n^{i,j}, n = 1, 2, \dots\}$ and (10) holds for the associated renewal processes.

ASSUMPTION C. For every state (q, z_1, z_2), every $n > 0$, we have $\tau \equiv \inf\{t : \|\mathbf{Q}(t)\| \ge n \mid \mathbf{Q}(0) = (q, z_1, z_2)\} < \infty$ with probability one, under any scheduling policy.

One way to verify Assumption B is via the following sufficient condition for the large deviations bounds to hold for i.i.d. sequences.

LEMMA 1. *Suppose* $\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_m, \dots,$ *is a nonnegative i.i.d. sequence with* $\mathbb{E}[\mathbf{Z}_1] = \alpha$ *such that there exists a function* $F(\theta), \theta \ge 0$, *taking values in* $\Re_+ \cup \{\infty\}$, *which is finite on some interval* $[0, \theta_0]$ *and which satisfies*

$$\sup_{z \ge 0} \mathbb{E}\big[e^{\theta(Z_1 - z)} \mid Z_1 \ge z\big] \le F(\theta), \tag{11}$$

for every $\theta \ge 0$. *Then this sequence satisfies the LD bounds* (9) *and* (10).

The proof of Lemma 1 is provided in the Appendix. It is simple to check that condition (11) is satisfied by many distributions including the exponential, Erlang and any distribution with bounded support. Note that, by setting $z = 0$, condition (11) implies that the distribution of Z_1 has a moment generating function for $\theta \in [0, \theta_0]$.

Assumption C is intentionally broad, in that it does not involve the stochastic primitives directly. The assumption holds for a wide range of distributions, given Assumptions A and B. For example, if at least one of the service time distributions has unbounded support, then Assumption C holds. However, Assumption C holds under even weaker conditions.

We adopt Assumptions A, B and C for the remainder of the paper. Whenever we talk about the probability $\mathbb{P}\{\cdot\}$ of any event, the probability is understood with respect to the stochastic processes $\{\mathbf{X}_n^i, \mathbf{S}_n^{i,j}\}$. If the vector of initial queues $\mathbf{Q}(0)$ is a random vector itself, then the probability is also with respect to the probability distribution of $\mathbf{Q}(0)$.

2.1.3. *Stability and rate stability.* One of the main features one desires to have in a multitype queueing network is stability. Various equivalent definitions of stability have been used in the literature, among which positive Harris recurrence is one of the most commonly used definitions. Under the condition that the interarrival times $\{X_k^i\}$ are unbounded and spread out (see [7]), then positive Harris recurrence is defined as follows.

DEFINITION 2. A multitype queueing network operating under a scheduling policy $\mathcal{U}$ is defined to be Harris recurrent if there exists $b > 0$ such that for any initial vector of queue lengths $\mathbf{Q}(0)$, the time $\tau = \inf\{t : \|Q(t)\| \le b\}$ is finite with probability one. The network is defined to be positive Harris recurrent or stable, if, in addition, $\mathbb{E}[\tau] < \infty$, where the expectation is conditioned on the initial vector of queue lengths $\mathbf{Q}(0)$. The network is defined to be globally stable if it is stable for every nonidling scheduling policy $\mathcal{U}$.

The positive Harris recurrence property, under some additional technical assumptions, implies the existence of a unique stationary distribution for the queue length process $\mathbf{Q}(t)$.

A somewhat weaker definition of stability is rate stability. This is the form of stability we are primarily concerned with in this paper.

DEFINITION 3. A multitype queueing network operating under a scheduling policy $\mathcal{U}$ is defined to be rate stable if for every type i, $\lim_{t\to\infty} \frac{\mathbf{D}_{i,J_i}(t)}{t} = \lambda_i$, a.s. The network is defined to be globally rate stable if it is rate stable for every nonidling scheduling policy $\mathcal{U}$.

In words, rate stability means with probability one the arrival rate is equal to the departure rate. From (1) and (2), rate stability implies $\lim_{t\to\infty} \frac{\mathbf{D}_{i,j}(t)}{t} = \lambda_i$ and $\lim_{t\to\infty} \frac{\mathbf{Q}_{i,j}(t)}{t} = 0$ a.s. for all i, j. In other words, for a rate stable system, even if the total queue length $\|\mathbf{Q}(t)\|$ diverges as t goes to infinity, it grows, at most, at a sub-linear rate a.s.

2.2. *Fluid model.*

2.2.1. *Fluid equations.* Fluid models are continuous deterministic counterparts of stochastic queueing networks, intended to capture the most essential dynamic properties of the queue length process. The term fluid model is sometimes used interchangeably with the terms “fluid limits” and “functional law of large numbers.” For many types of queueing networks (see, e.g., [2, 6, 7, 13, 20]), it has been established that the rescaled queue length process $\mathbf{Q}(nt)/n$ for a large scaling parameter n converges weakly to a certain continuous deterministic process, satisfying a series of functional equations, which we describe below. To avoid confusion, we define the *fluid limit model* to be the set of weak limits of $\mathbf{Q}(nt)/n$ as $n \to \infty$, and we define the fluid model to be the set of solutions of the system of equations below (formal definition follows). Then the set of fluid limits is a subset of the set of solutions to the fluid model.

Given a multitype queueing network with arrival rates λ_i and service rates $\mu_{i,j}$, the corresponding HOL fluid model (or fluid network) is defined by the following system of equations and inequalities with time dependent variables

$\bar{Q}_{i,j}(t), \bar{A}_i(t), \bar{D}_{i,j}(t), \bar{T}_{i,j}(t), t \ge 0$. We first provide the system of equations and inequalities, and immediately after we give a physical explanation for each of these equations. For every $i = 1, \dots, I$, $j = 1, 2, \dots, J_i$, $\sigma = \sigma_1, \dots, \sigma_J$, $t \ge 0$ and $0 \le t_1 \le t_2$,

$$\bar{Q}_{i,1}(t) = \bar{Q}_{i,1}(0) + \bar{A}_i(t) - \bar{D}_{i,1}(t), \tag{12}$$

$$\bar{Q}_{i,j}(t) = \bar{Q}_{i,j}(0) + \bar{D}_{i,j-1}(t) - \bar{D}_{i,j}(t), \tag{13}$$

$$\bar{A}_i(t) = \lambda_i t, \tag{14}$$

$$\bar{D}_{i,j}(t) = \mu_{i,j} \bar{T}_{i,j}(t), \tag{15}$$

$$\sum_{(i,j)\in\sigma} \bigl(\bar{T}_{i,j}(t_2) - \bar{T}_{i,j}(t_1)\bigr) \le t_2 - t_1, \tag{16}$$

$$\bar{Q}_{i,j}(t), \bar{A}_i(t), \bar{D}_{i,j}(t), \bar{T}_{i,j}(t) \in \Re_+. \tag{17}$$

In addition, for all $i, j, \bar{T}_{i,j}(t)$ is a nondecreasing function of t and $\bar{A}_{i,j}(0)$, $\bar{T}_{i,j}(0) = 0$.

The value of $\bar{Q}_{i,j}(t)$ represents the total amount of fluid present in buffer (i, j) at time t. We also refer to it as class (i, j) fluid. $\bar{A}_i(t)$ represents the total amount of fluid corresponding to type i, that arrived externally during the time interval $[0, t]$. The fluid arrival process is assumed to be linear with rate λ_i, hence, (14). $\bar{D}_{i,j}(t)$ is the amount of class (i, j) fluid that was processed by station $\sigma(i, j)$ during $[0, t]$. $\bar{T}_{i,j}(t)$ represents the portion of the time interval $[0, t]$ that station $\sigma(i, j)$ spent processing class (i, j) fluid. Inequality (16) enforces the physical constraint that any given station can spend at most 100% of its time processing fluid.

Equations (12) and (13) are simply flow conservation equations: all class $(i, j-1)$ fluid becomes class (i, j) fluid after processing, for all $j \le J_i$, and class (i, J_i) fluid leaves the network after processing. The last constraint (17) simply says that all the variables involved are nonnegative real numbers. Note that only the expectations $1/\lambda_i = \mathbb{E}[\mathbf{X}_1^i]$ and $1/\mu_{i,j} = \mathbb{E}[\mathbf{S}_1^{i,j}]$ of interarrival and service times appear in the fluid model. The higher-order moments of the network primitives are not reflected in the model.

For each station σ, we let

$$\bar{Q}_\sigma(t) = \sum_{(i,j)\in\sigma} \bar{Q}_{i,j}(t), \tag{18}$$

that is, $\bar{Q}_\sigma(t)$ is the total fluid level in station σ at time t. Also let

$$\bar{T}_\sigma(t) = \sum_{(i,j)\in\sigma} \bar{T}_{i,j}(t). \tag{19}$$

So, $\bar{T}_\sigma(t)$ is the total amount of time station σ spent processing fluid during the time interval $[0, t]$. Equivalently, $\bar{I}_\sigma(t) \equiv t - \bar{T}_\sigma(t)$ represents the cumulative amount of idling experienced by station σ during the time interval $[0, t]$.

From (16), it follows that the function $\bar{T}_{i,j}(t)$ is Lipschitz continuous. Using (12)–(17), it can be checked that all of $\bar{Q}_{i,j}(t), \bar{A}_i(t), \bar{D}_{i,j}(t), \bar{T}_{i,j}(t)$ are also Lipschitz continuous. Any solution $(\bar{Q}_{i,j}(t), \bar{A}_i(t), \bar{D}_{i,j}(t), \bar{T}_{i,j}(t))$ of the system of equations and inequalities (12)–(17) is defined to be a fluid solution. For simplicity, henceforth we use $(\bar{Q}(t), \bar{T}(t))$ to denote a fluid solution, where $\bar{Q}(t)$ and $\bar{T}(t)$ stand, respectively, for vectors $(\bar{Q}_{i,j}(t))$ and $(\bar{T}_{i,j}(t))$. A fluid solution $(\bar{Q}(t), \bar{T}(t))$ is defined to be nonidling if for every station σ, $\bar{I}_\sigma(t)$ increases only at times t when $\bar{Q}_\sigma(t) = 0$. Formally, the fluid solution is nonidling if for every station σ,

$$\int_0^\infty \bar{Q}_\sigma(t)\, d\bar{I}_\sigma(t) = 0. \tag{20}$$

The integral is well defined because $\bar{I}_\sigma(t)$ is a Lipschitz continuous function and, as a result, is almost everywhere differentiable in $\Re_+$ with respect to the Lebesgue measure on $\Re_+$.

DEFINITION 4. The set of nonidling feasible solutions to the system of equations (12)–(17) and (20) is defined to be the nonidling *fluid model*.

When a queueing network operates under a specific scheduling policy, for example, under a fixed buffer priority policy, additional constraints can be added to the fluid equations in order to reflect the policy. In this paper we are only considering the case of all the nonidling policies, and thus the nonidling fluid model defined is the one of interest. For the remainder of the paper, we drop the modifier "nonidling" and simply refer to the "fluid model."

The following lemmas are used later in the paper. The proofs of both lemmas are straightforward and thus omitted (note that Lemma 3 appeared as Property 2 in [4]).

LEMMA 2. *Suppose* $(\bar{Q}(t), \bar{T}(t))$ *is a fluid solution defined over a time interval* $[0, \theta]$. *Then* $\bar{Q}'(t) \equiv \bar{Q}(0) + \frac{t}{\theta}(\bar{Q}(\theta) - \bar{Q}(0)), \bar{T}'(t) \equiv \frac{t}{\theta}\bar{T}(\theta)$ *defined over* $[0, \theta]$ *is also a fluid solution. Moreover, suppose the solution* $(\bar{Q}(t), \bar{T}(t))$ *is nonidling and for every station* σ, *either* $\bar{Q}_\sigma(t) > 0$ *for all* $t \in [0, \theta]$ *or* $\bar{Q}_\sigma(0) = \bar{Q}_\sigma(\theta) = 0$. *Then the solution* $(\bar{Q}'(t), \bar{T}'(t))$ *is also nonidling.*

LEMMA 3. *Suppose* $(\bar{Q}(t), \bar{T}(t))$ *is a nonidling fluid solution defined over a time interval* $[0, \theta]$. *Then for any* $\beta > 0$, $\bar{Q}'(t) \equiv \beta\bar{Q}(\beta^{-1}t)$, $\bar{T}'(t) \equiv \beta\bar{T}(\beta^{-1}t)$ *is a nonidling fluid solution defined over the interval* $[0, \beta\theta]$.

In the proofs in later sections, we need to define certain types of fluid models with a finite decomposition property. We define this notion below.

DEFINITION 5. A fluid model is defined to satisfy the *Finite Decomposition Property* (*FDP*) if there exist values $\nu, B > 0$, with the following property. For every nonidling fluid solution $(\bar{Q}(t), \bar{T}(t))$ defined over an interval $[0, \theta]$ such that $\bar{Q}(t) \neq 0$ on this interval, there exist a nonidling fluid solution $(\tilde{Q}(t), \tilde{T}(t))$ also defined over $[0, \theta]$ and a sequence of times instances $0 = t_0 < t_1 < t_2 < \cdots < t_M = \theta$ such that:

1. $M \leq \nu\theta \sup_{0\leq t\leq\theta} \frac{1}{\|\bar{Q}(t)\|} + B$ and $\inf_{0\leq t\leq\theta} \|\tilde{Q}(t)\| \geq \inf_{0\leq t\leq\theta} \|\bar{Q}(t)\|$.
2. $\tilde{Q}(t_m) = \bar{Q}(t_m)$ for all $m = 0, 1, \dots, M$.
3. For each interval (t_r, t_{r+1}), $0 \leq r \leq M - 1$ and each station σ either $\tilde{Q}_\sigma(t) > 0$ for all $t \in (t_r, t_{r+1})$ or $\tilde{Q}_\sigma(t) = 0$ for all $t \in (t_r, t_{r+1})$.

The next proposition shows that the FDP requirement is not restrictive for fluid models arising from two station networks.

PROPOSITION 1. *Fluid networks with two stations* $(J = 2)$ *satisfy FDP.*

Although we only consider multitype fluid networks in this paper, the proposition actually holds for any two station fluid network, for example, networks with proportional routing. This general form of Proposition 1 is proved in Section 4.2. At this point we do not know whether FDP holds for general networks (i.e., with $J > 2$).

2.2.2. *Global stability and global weak stability.* Just as for stochastic queueing networks, we can define stability and global stability for fluid networks.

DEFINITION 6. A fluid solution $(\bar{Q}(t), \bar{T}(t))$ is defined to be stable if there exists a $\tau < \infty$ such that $\bar{Q}(t) = 0$ for all $t \geq \tau$. A fluid model is defined to be globally stable if there exists a $\tau < \infty$ such that every nonidling fluid solution $(\bar{Q}(t), \bar{T}(t))$ satisfying $\|\bar{Q}(0)\| = 1$ also satisfies $\bar{Q}(t) = 0$ for all $t \geq \tau$.

REMARKS. 1. The condition $\|\bar{Q}(0)\| = 1$ in the definition above is a necessary scaling condition. One cannot have a uniform emptying time τ without a bound on the initial state.

2. The definition of global stability is somewhat different from the perhaps more natural: “network is defined to be globally stable if it is stable for all nonidling policies.” While it is possible that both definitions are equivalent and it is known to hold in many cases, it has not yet been established in general. Definition 6 is used more often because it simplifies certain technical considerations.

Below, we define a stability notion for fluid networks which is the analogue of the rate stability definition for stochastic networks.

DEFINITION 7. A fluid model is defined to be globally weakly stable if for any nonidling fluid solution $(\bar{Q}(t), \bar{T}(t))$, $\bar{Q}(0) = 0$ implies $\bar{Q}(t) = 0$ for all $t \geq 0$.

In words, a fluid model is weakly stable if one cannot construct a nonzero fluid solution which starts from zero. We did not introduce the notion of a weakly stable fluid solution, since this would just mean introducing a trivial $\bar{Q}(t) = 0$ solution [also it is easy to check that $\bar{Q}(t) = 0$, for all t implies $T_{i,j}(t) = \frac{\lambda_i}{\mu_i} t$ for all (i, j)].

2.3. *The connections between stochastic and fluid queueing networks.* The most immediate connection between a stochastic network and the corresponding fluid queueing network is provided by the results of Dai [7] and Stolyar [21]. Roughly speaking, they show that for a broad class of scheduling policies, if a stochastic network is operating under a policy $\mathcal{U}$, each weak limit $\bar{Q}(t) = \lim_n \frac{\mathbf{Q}(nt)}{t}$ and $\bar{T}(t) = \lim_n \frac{\mathbf{T}(nt)}{t}$ of the stochastic queue length process $\mathbf{Q}(t)$ and cumulative work process $\mathbf{T}(t)$, with a sequence of initial states $\mathbf{Q}(0) = \lfloor n\gamma \rfloor$, where γ is a fixed positive constant, is a deterministic continuous function $(\bar{Q}(t), \bar{T}(t))$ which is a fluid solution of the corresponding fluid model. If the policy $\mathcal{U}$ is nonidling, then each obtained fluid solution is also nonidling. Thus, the queue length process, after an appropriate rescaling using certain scaled initial states, converges to a fluid solution.

This rescaling process provides the basic tool for connecting the stability of stochastic and fluid networks. In fact, this connection was the primary motivation for introducing fluid model techniques [20]. The following theorem establishes a fundamental relationship between the stability of the stochastic and fluid models.

THEOREM 4 (Dai [7], Stolyar [21]). *Consider a multitype queueing network. If the corresponding fluid model is globally stable, then the stochastic network is globally stable.*

Theorem 4 actually holds for a broader class of networks and also for networks operating under specific scheduling policies. If one is given a particular scheduling policy $\mathcal{U}$, one can sometimes identify additional constraints that the fluid limits $\lim_n \mathbf{Q}(nt)/n$ must satisfy.

More relevant to the topic of the present paper is the following related result.

THEOREM 5 (Chen [4]). *Consider a multitype queueing network. If the corresponding fluid model is globally weakly stable, then the stochastic network is globally rate stable.*

Our understanding of global stability and global weak stability is fairly complete for fluid models corresponding to queueing networks with two stations ($J = 2$), thanks to the results of Bertsimas, Gamarnik and Tsitsiklis [1] and Dai and

VandeVate [12]. Both of these papers obtain necessary and sufficient conditions for global stability of fluid networks for the case $J = 2$. Moreover, a certain parameter ρ^* is introduced in [12]. This parameter is called the maximum virtual traffic intensity. It is shown that the fluid model is globally stable iff $\rho^* < 1$ and is globally weakly stable iff $\rho^* \le 1$. The condition $\rho^* \le 1$ then implies rate stability of the underlying stochastic network by Theorem 5. One of the main results of our paper is to establish a converse: $\rho^* > 1$ implies the stochastic network is not globally rate stable. In particular, $\rho^* \le 1$ is the tight global rate stability condition for multitype networks with two stations.

3. Main results.

In this section we provide the main results and corollaries of this paper. All proofs, along with the needed lemmas, are provided in Section 4. Our first result concerns the structural properties of nonidling fluid solutions. The result is introduced primarily because it is needed to prove the main result of the paper, but we believe that it is interesting in its own right and thus state the result in this section.

THEOREM 6. *Suppose the fluid model of a multitype queueing network is not weakly stable. Then there exists a positive constant* $\gamma > 0$ *such that for any initial state* $q \in \Re^d_+$, *there exists a nonidling fluid solution* $(\bar{Q}(t), \bar{T}(t))$ *satisfying* $\bar{Q}(0) = q$ *and* $\|\bar{Q}(t)\| \ge \gamma t$ *for all* $t \ge 0$. *Namely, the solution is linearly divergent. Moreover, this solution satisfies*

$$\inf_{t\ge 0} \|\bar{Q}(t)\| \ge \frac{\|q\|}{2} \min\left(\frac{\gamma}{C}, 1\right), \tag{21}$$

where C is defined by (8).

Intuitively, the notion of a fluid model not being weakly stable seems weaker than linear divergence. In particular, a fluid model is not weakly stable if there exists a solution which “pops up from zero” at some point, after starting in the zero state. Theorem 6 shows that if one solution pops up, then a different solution can be constructed which diverges to infinity linearly, that is, we construct a stronger fluid solution (in the sense of instability) from a seemingly weaker solution. This stronger fluid solution can then be used to infer the instability of a class of associated stochastic networks. Finally, we note that the divergent solution can be constructed from any initial state $\bar{Q}(0) = q \in \Re^d_+$, including the zero state.

We are now prepared to state the main result of the paper, which connects the instability of fluid models and stochastic networks.

THEOREM 7. *Consider a multitype stochastic network satisfying Assumptions* A, B *and* C. *Suppose the associated fluid model is not globally weakly stable, and satisfies FDP. Then, for any initial state* $(q, z_1, z_2) \in \mathbb{Z}^d_+ \times \Re^{I+d}$, *there exists a nonidling scheduling policy for which the resulting queue level process satisfies*

$\liminf_{t\to\infty} \frac{\|\mathbf{Q}(t)\|}{t} > 0$ *with probability one. In particular, the stochastic process associated with the queueing network is unstable, under some nonidling policy.*

The rate of divergence to infinity implied by the theorem above will be explicit. We will show that constructed policy results in

$$\liminf_{t\to\infty} \frac{\|\mathbf{Q}(t)\|}{t} \geq \frac{\max(\gamma/C, 1)}{8\max(1, 3/\gamma)} \tag{22}$$

with probability one. The import of Theorem 7 is more apparent from the corollaries provided below.

COROLLARY 1. *Consider a multitype stochastic network with $J = 2$. If the associated fluid model is not globally weakly stable, then the queueing network is unstable in the sense that* $\liminf_{t\to\infty} \frac{\|\mathbf{Q}(t)\|}{t} > 0$ *with probability one from each initial state under some nonidling scheduling policy.*

Corollary 1 follows from Theorem 7 and Proposition 1, which states that FDP holds for fluid networks with two stations. Recall that one motivation for our work is the stability Theorems 4 and 5. Thus, Corollary 1 provides a complete converse of Theorem 5 for two station multitype networks. A missing piece in the theory for general J is to determine if all fluid models satisfy FDP. If such a result holds, then Theorem 7 would imply a converse for networks with an arbitrary number of stations.

We note also that Theorem 5 is valid when we consider fluid and queueing networks under specific scheduling policies. However, for networks operating under specific policies (rather than a class of policies), a general converse to the theorems of Chen and Dai is not possible as demonstrated in [10].

Dai and VandeVate [12] derived explicit necessary and sufficient conditions for global weak stability of fluid models of multitype networks in terms of a certain parameter ρ^* related to the so-called virtual traffic intensity and push start conditions. They prove that such fluid networks are weakly stable iff $\rho^* \leq 1$. Considering Theorem 5 along with Corollary 1, those results now yield complete necessary and sufficient conditions for rate stability of two station stochastic mutlitype networks.

COROLLARY 2. *A stochastic two station multitype network is globally rate stable if and only if $\rho^* \leq 1$.*

4. Proofs of main results.

In this section we provide all of the proofs of our main results. The first proof, of Theorem 6, shows that if the fluid is not globally weakly stable, there exists a linearly divergent fluid solution.

4.1. *Linearly divergent fluid solutions.*

PROOF OF THEOREM 6. We assume in the theorem that a given fluid model is not globally weakly stable. Hence, there exists a nonidling solution which satisfies $\bar{Q}(0) = 0$ and $\bar{Q}(t_0) \neq 0$ for some $t_0 > 0$. First note that, without the loss of generality, we may assume that $\bar{Q}(t) \neq 0$ for all $0 < t \leq t_0$. Otherwise, we can find $\hat{t} = \sup\{0 \leq t < t_0 : \bar{Q}(t) = 0\}$ and consider the fluid solution on $[\hat{t}, t_0]$ only. Note that $\hat{t} < t_0$ by the continuity of $\bar{Q}(t)$ and the fact that $\bar{Q}(t_0) \neq 0$. Next, using Lemma 3 with some $\beta > 0$, we can obtain a new solution defined on $[0, \beta t_0]$ with $\bar{Q}'(0) = 0$ and $\|\bar{Q}'(\beta t_0)\| = \beta \|\bar{Q}(t_0)\|$. If we set $\beta = t_0^{-1}$, then we have a solution defined on $[0, 1]$ with $\bar{Q}'(0) = 0$ and $\|\bar{Q}'(1)\| = \|\bar{Q}(t_0)\|/t_0$. Hence, again without loss of generality, we set $t_0 = 1$, that is, we assume we are given a nonidling solution with $\bar{Q}(0) = 0$ and $\bar{Q}(1) \neq 0$.

We now build a new fluid solution by constructing it iteratively over the intervals $[0, 1), [1, 2), [2, 4), \dots, [2^n, 2^{n+1}), \dots$. We denote the solution that is constructed in this manner by $(\bar{Q}^o(t), \bar{T}^o(t))$. For the initial interval $[0, 1)$, consider our initial fluid solution $\bar{Q}(t)$ satisfying $\bar{Q}(0) = 0$, $\bar{Q}(1) \neq 0$. We first modify the solution by setting $\bar{Q}(0) = q$, where $q \in \Re^d_+$. Next, for every $t \leq 1$ and every class (i, j), on the interval $[0, t]$ we spend exactly $\bar{T}_{i,j}(t)$ time units processing class (i, j) flow, *plus* whatever necessary additional amount is required to make the solution nonidling. In other words, we can think of the flow "created" and "processed" by the nonweakly stable solution $\bar{Q}(t)$ as high priority flow, and the remaining flow as low priority flow. Note that the allocation of the additional processing effort required is not necessarily uniquely determined by the original allocation $\bar{T}(t)$. In any case, the resulting solution satisfies $\bar{Q}^o(0) = q$ and $\bar{Q}^o_{i,j}(t) \geq \bar{Q}_{i,j}(t)$ for all classes (i, j) and $t \leq 1$. In particular, $\|\bar{Q}^o(1)\| \geq \|\bar{Q}(1)\| > 0$.

Assume now the solution has been constructed over the time horizon $[0, 2^n]$ for $n \geq 0$. We now extend it over $[2^n, 2^{n+1}]$. The idea of the construction is similar to the first interval, except that we "stretch" the original solution $\bar{Q}(t)$ by a factor of 2^n and then use this solution to extend our current solution by defining it on $[2^n, 2^{n+1}]$. That is, consider the scaled solution $(\beta \bar{Q}(\beta^{-1}t), \beta \bar{T}(\beta^{-1}t))$ with $\beta = 2^n$. This solution is defined over $t \in [0, 2^n)$. Next, for each $t \in [2^n, 2^{n+1}]$, let $\bar{T}^o(t)$ be defined by $\bar{T}^o(t) - \bar{T}^o(2^n) = 2^n \bar{T}(2^{-n}(t - 2^n))$, plus any extra processing effort required to make the solution nonidling.

It can be easily checked that the resulting solution $\bar{Q}^o(t)$ satisfies $\bar{Q}^o_{i,j}(t) \geq 2^n \bar{Q}_{i,j}(2^{-n}(t - 2^n))$ for all $t \in [2^n, 2^{n+1}]$ and all i, j which implies $\|\bar{Q}^o(t)\| \geq 2^n \|\bar{Q}(2^{-n}(t - 2^n))\|$. In particular, $\|\bar{Q}^o(2^{n+1})\| \geq 2^n \|\bar{Q}(1)\|$.

We have constructed a nonidling fluid solution $\bar{Q}^o(t)$ which diverges to infinity at time instances $t_n = 2^n, n = 0, 1, \dots$. To complete the proof of the theorem, we show that, for some constant $\gamma_0 > 0$, $\|\bar{Q}^o(t)\| \geq \gamma_0 2^n$ for all $t \in [2^n, 2^{n+1}]$. First let us show that this implies the theorem. For any $t > 0$, find the largest integer n

such that $2^n \leq t$, that is, let $n = \lfloor \log_2 t \rfloor$. We have $\|\bar{Q}^o(t)\| \geq \gamma_0 2^n \geq \gamma_0 2^{\log_2 t - 1} = \gamma_0 t/2$. Setting $\gamma = \gamma_0/2$, we obtain the result.

To show the existence of γ_0, note that for any $t_1 < t_2$ and any feasible fluid solution $\bar{Q}(\cdot)$, we have

$$\begin{aligned} \|\bar{Q}(t_2)\| &\geq \|\bar{Q}(t_1)\| - \sum_{1 \leq i \leq I} \mu_{i,J_i}(t_2 - t_1) \\ &\geq \|\bar{Q}(t_1)\| - C(t_2 - t_1). \end{aligned} \tag{23}$$

This implies that, for all $t \in [2^n, 2^n + 2^{n-1}\|\bar{Q}(1)\|/(2C)]$, $\bar{Q}^o(t)$ satisfies

$$\begin{aligned} \|\bar{Q}^o(t)\| &\geq \|\bar{Q}^o(2^n)\| - C(t - 2^n) \\ &\geq 2^{n-1}\|\bar{Q}(1)\| - C(t - 2^n) \\ &\geq 2^{n-2}\|\bar{Q}(1)\|. \end{aligned}$$

If $2^n + 2^{n-1}\|\bar{Q}(1)\|/(2C) \geq 2^{n+1}$, then we simply set $\gamma_0 = (1/4)\|\bar{Q}(1)\|$. Otherwise, let

$$\gamma_1 = \min\left\{\|\bar{Q}(t)\| : \frac{\|\bar{Q}(1)\|}{4C} \leq t \leq 1\right\}.$$

This minimum exists since $\bar{Q}(t)$ is continuous and it is positive since $\|\bar{Q}(t)\| > 0$ for all $0 < t \leq 1$. Then, for all $2^n + 2^{n-1}\|\bar{Q}(1)\|/(2C) \leq t \leq 2^{n+1}$, we have $\|\bar{Q}^o(t)\| \geq 2^n\|\bar{Q}(2^{-n}(t - 2^n))\| \geq 2^n\gamma_1$. We take $\gamma_0 = \min\{(1/4)\|\bar{Q}(1)\|, \gamma_1\}$ and we have proven the first inequality in the theorem statement.

The last part of the proposition follows almost immediately. Using (23) with $t_1 = 0$ and $t_2 = t$, we have $\|\bar{Q}(t)\| \geq \|q\| - Ct \geq \|q\|/2$ for $t \leq \|q\|/(2C)$. On the other hand, by construction, $\|\bar{Q}(t)\| \geq \gamma t \geq \gamma\|q\|/(2C)$, whenever $t \geq \|q\|/(2C)$. This completes the proof of the theorem. □

Theorem 6 will be used for proving our main result, Theorem 7. Specifically, we will construct a nonidling scheduling policy for the discrete network which, with high probability, results in a trajectory very close to the fluid trajectory built in the proof of Theorem 6. We will use the large deviations bounds (9) and (10) multiple times to obtain bounds on the deviation between the fluid and stochastic trajectories.

4.2. *FDP in fluid networks with two stations.*

PROOF OF PROPOSITION 1. Consider a network with two stations, σ_1 and σ_2, and suppose we have a nonidling fluid solution $(\bar{Q}(t), \bar{T}(t))$ which is nonzero over time interval $[0, \theta]$. By continuity, $\inf_{0 \leq t \leq \theta}\|\bar{Q}(t)\| > 0$. The next result follows from Proposition 1 in [1]. There exists a nondecreasing sequence t_i such that $\sup_i t_i = \theta$ and such that for all times less than θ the following hold:

- $\bar{Q}_{\sigma_1}(t_{4m+1}) > 0$, $\bar{Q}_{\sigma_2}(t_{4m+1}) = 0$ and for $t \in [t_{4m+1}, t_{4m+2}]$, $\bar{Q}_{\sigma_1}(t) > 0$;
- $\bar{Q}_{\sigma_1}(t_{4m+2}) > 0$, $\bar{Q}_{\sigma_2}(t_{4m+2}) = 0$ and for $t \in (t_{4m+2}, t_{4m+3})$, $\bar{Q}_{\sigma_1}(t)$, $\bar{Q}_{\sigma_2}(t) > 0$;
- $\bar{Q}_{\sigma_2}(t_{4m+3}) > 0$, $\bar{Q}_{\sigma_1}(t_{4m+3}) = 0$ and for $t \in [t_{4m+3}, t_{4m+4}]$, $\bar{Q}_{\sigma_2}(t) > 0$;
- $\bar{Q}_{\sigma_2}(t_{4m+4}) > 0$, $\bar{Q}_{\sigma_1}(t_{4m+4}) = 0$ and for $t \in (t_{4m+4}, t_{4m+5})$, $\bar{Q}_{\sigma_1}(t)$, $\bar{Q}_{\sigma_2}(t) > 0$.

Moreover, one of $t_i, i = 1, 2, 3, 4$, is equal to zero. When t_2, t_3 or t_4 is zero, t_i with lower value of i is not defined.

The characterization above essentially divides the trajectory of a fluid solution into four different segments. On the segment of the trajectory between t_{4m+1} and t_{4m+2}, the trajectory is either on the boundary of the state space [where $\bar{Q}_{\sigma_2}(t) = 0$] or in the interior of the state space. We next claim that such a segment can be "linearized" such that it remains a nonidling solution, yet $\bar{Q}_{\sigma_2}(t) = 0$ for all $t \in [t_{4m+1}, t_{4m+2}]$. In other words, the linearized solution is on the boundary for the entire interval. To achieve the linearization, we define

$$\tilde{Q}(t) = \bar{Q}(t_{4m+1}) + \frac{t - t_{4m+1}}{t_{4m+2} - t_{4m+1}}[Q(t_{4m+2}) - Q(t_{4m+1})]$$

and

$$\tilde{T}(t) = \bar{T}(t_{4m+1}) + \frac{t - t_{4m+1}}{t_{4m+2} - t_{4m+1}}[T(t_{4m+2}) - T(t_{4m+1})],$$

for all $t \in [t_{4m+1}, t_{4m+2}]$. Using Lemma 2, it follows that the new solution $(\tilde{Q}(t), \tilde{T}(t))$ is both feasible and nonidling, given that the original solution was also. In a similar manner, we linearize the fluid solution $(\bar{Q}(t), \bar{T}(t))$ on all intervals of the form $[t_{4m+3}, t_{4m+4}]$. Hence, in each interval the new solution remains on one of the axes, unless it is crossing the interior, from one axis to the other.

We now demonstrate that $(\tilde{Q}(t), \tilde{T}(t))$ has the properties described in Definition 5. First, we claim that for each m,

$$\begin{aligned} t_{4m+3} - t_{4m+1} &\geq \inf_{0 \leq t \leq \theta} \|\bar{Q}(t)\|/C, \\ t_{4m+3} - t_{4(m+1)+1} &\geq \inf_{0 \leq t \leq \theta} \|\bar{Q}(t)\|/C. \end{aligned} \tag{24}$$

Indeed, by construction,

$$\tilde{Q}_{\sigma_1}(t_{4m+1}) > 0, \qquad \tilde{Q}_{\sigma_2}(t_{4m+1}) = 0$$

and

$$\tilde{Q}_{\sigma_2}(t_{4m+3}) > 0, \qquad \tilde{Q}_{\sigma_1}(t_{4m+3}) = 0.$$

In particular,

$$\tilde{Q}_{\sigma_1}(t_{4m+1}) = \|\tilde{Q}(t_{4m+1})\| \geq \inf_{0 \leq t \leq \theta} \|\tilde{Q}(t)\| \geq \inf_{0 \leq t \leq \theta} \|\bar{Q}(t)\| > 0.$$

Note that total rate at which fluid can depart from a given station is bounded above by $\sum_{i,j}\mu_{i,j} < C$. Thus, since $\tilde{Q}_{\sigma_1}(t_{4m+3}) = 0$, we have $t_{4m+3} - t_{4m+1} \geq \inf_{0\leq t\leq\theta}\|\bar{Q}(t)\|/C$. An analogous argument demonstrates that $t_{4m+3} - t_{4(m+1)+1} \geq \inf_{0\leq t\leq\theta}\|\bar{Q}(t)\|/C$. Since the interval lengths are bounded strictly away from zero, the total number of points t_i in $[0,\theta]$ is at most $(2\theta C/\inf_{0\leq t\leq\theta}\|\bar{Q}(t)\|) + 2$, where the $+2$ accounts for the end points of $[0,\theta]$. Setting $\nu = 2C$, $B = 2$ yields the first FDP property (1). Properties (2) and (3) are automatically satisfied by our construction of $(\tilde{Q}(t), \tilde{T}(t))$ above. □

4.3. *Transient paths in the stochastic network.* Most of this section is devoted to the proof of Theorem 14 of Section 4.4, which, as we will show, implies the main result of our paper, Theorem 7. In the proof we repeatedly use probabilistic bounds of the form $c_1\exp(-c_2 n)$, where $c_1, c_2 > 0$ are constants which depend on the parameters of our queueing network and n is a scaling parameter which takes on a large value. In various expressions, c_2 is usually related to the constant L appearing in the large deviations bounds in (9) and (10) and the network parameters $\lambda_i, \mu_{i,j}, |I|, C$, as well as parameter γ introduced in Theorem 6. We will also be considering finite sums of the bounds of the form $c'_1\exp(-c_1 n) + c'_2\exp(-c_2 n) + \cdots + c'_m\exp(-c_m n)$. In general, the c_i, c'_i take on different values and m is a constant, independent of n. Such sums can be bounded above by $c'\exp(-cn)$ for $c = \min_{1\leq k\leq m} c_k$ and $c' = \sum c'_i$.

In our proofs, the actual values of the constants are not important, only the fact that they are independent of n. Therefore, to simplify the exposition, we simply use the notation $O(\exp(-\Theta(n)))$ and we write expressions like $O(\exp(-\Theta(n))) + O(\exp(-\Theta(n))) = O(\exp(-\Theta(n)))$, where the standard notation $O(\cdot)$ and $\Theta(\cdot)$ hides the actual constants c and c'.

4.3.1. *Proof preliminaries and the scheduling policy $\mathcal{U}$.* In order to precisely state the next series of detailed results, we need to define a nonidling policy $\mathcal{U}$. The definition of this policy involves a number of preliminary observations and definitions.

First, let

$$\theta = \max\left(1, \frac{3}{\gamma}\right). \tag{25}$$

The parameter θ depends only on parameters of the model since γ depends only on the parameters of the model.

Consider any initial state $(q, z_1, z_2) \in \mathbb{Z}^d_+ \times \Re^{I+d}_+$. Let $n = \|q\|$. By Theorem 6, there exists a nonidling fluid solution $(\bar{Q}(t), \bar{T}(t))$ which satisfies $\bar{Q}(0) = q$ and $\|\bar{Q}(t)\| \geq \gamma t$ for all $t \geq 0$. Since FDP is assumed then by Proposition 1, the solution $(\bar{Q}(t), \bar{T}(t))$ can be modified to a solution which satisfies properties described in Definition 5. Let

$$\theta_0 = \theta\|q\| = \|q\|\max\left(1, \frac{3}{\gamma}\right), \tag{26}$$

in which case we have

$$\|\bar{Q}(\theta_0)\| \geq 3\|q\|. \tag{27}$$

By Theorem 6, the fluid solution is also such that

$$\inf_t \|\bar{Q}(t)\| \geq \frac{\|q\|}{2} \min\left(\frac{\gamma}{C}, 1\right). \tag{28}$$

Since FDP is satisfied, there exists another solution $(\tilde{Q}(t), \tilde{T}(t))$ and a sequence $0 = s_0 < s_1 < \cdots < s_M = \theta_0$, such that $\inf_{0\leq t\leq\theta_0} \|\tilde{Q}(t)\| \geq \inf_{0\leq t\leq\theta_0} \|\bar{Q}(t)\|$ and, for every interval $[s_r, s_{r+1}]$ and for each station σ, either $\tilde{Q}_\sigma(t)$ is zero within (s_r, s_{r+1}), or it is strictly positive within (s_r, s_{r+1}). For simplicity, we assume that $(\bar{Q}(t), \bar{T}(t))$ is this modified solution. In such a modified solution we also note that

$$\begin{aligned} M &\leq \nu\theta_0 \sup_{0\leq t\leq\theta_0} \frac{1}{\|\tilde{Q}(t)\|} + B \leq \nu\theta_0 \frac{2}{\|q\|} \max\left(\frac{C}{\gamma}, 1\right) + B \\ &\leq 2\nu \max\left(1, \frac{3}{\gamma}\right) \max\left(\frac{C}{\gamma}, 1\right) + B, \end{aligned} \tag{29}$$

where we used (26) and (28). In particular, we obtain a bound on M which depends only on the parameters of the model (and is independent of $\|q\|$), since ν, γ and C depend only on the parameters of the model. Note, on the other hand, that the partition $s_r, r = 0, 1, \ldots, M$, does depend on q. Recalling the notation $\|q\| = n$, we rewrite (27) and (28) as

$$\|\bar{Q}(n\theta)\| \geq 3\|q\| = 3n \tag{30}$$

and

$$\inf_{0\leq t\leq\theta n} \|\bar{Q}(t)\| \geq \frac{n}{2} \min\left(\frac{\gamma}{C}, 1\right). \tag{31}$$

Our next goal is to describe a nonidling scheduling policy $\mathcal{U} = \mathcal{U}(\delta)$ implemented over the time horizon $[0, \theta_0] = [0, \theta n]$. Recall that our starting state is (q, z_1, z_2). In particular, $\mathbf{Q}(0) = \bar{Q}(0) = q$. The policy $\mathcal{U}$ attempts to mimic the fluid solution described above, over the same time interval. We parameterize the policy with a constant $\delta > 0$, which is any constant satisfying

$$\delta \leq \frac{1}{12C^{M+3}} \min\left(\frac{\gamma}{C}, 1\right). \tag{32}$$

Let $t_m = m\delta n$ for $m = 0, 1, \ldots, \lceil\theta/\delta\rceil$. We describe the policy $\mathcal{U}$ on each time interval $I_m = [t_m, t_{m+1})$. For each time interval I_m, each station σ nominally allocates $\bar{T}_{i,j}(t_m, t_{m+1})$ time units to serving class (i, j), for every class $(i, j) \in \sigma$. To be precise, we first order all the classes at a station in a fixed, but arbitrary manner. During the interval a class (i, j) is chosen for service, and we work on jobs from that class for $\bar{T}_{i,j}(t_m, t_{m+1})$ time units or until we exhaust the jobs from class (i, j). Note that we cannot reach the end of the interval I_m, by the feasibility of $\bar{T}(\cdot)$ over this interval. When we are done processing jobs of type (i, j), the next class in the chosen order of service is picked for processing. Note that we assume a preemptive resume mechanism when switching between classes. If after going through all the classes, the time spent is strictly less than $t_{m+1} - t_m$ and there are still jobs at the station, the station works on any available jobs. If no jobs are available, the station idles. Once the next time instance t_{m+1} occurs, the policy is "reset," in terms of the time allocations.

In other words, according to our scheduling policy, on each interval I_m each station tries to spend exactly the same amount of time on jobs in each class (i, j) as the fluid solution $(\bar{Q}(t), \bar{T}(t))$ does, while maintaining the nonidling requirement. Our main goal is to show that, in general, the resulting stochastic process stays fairly close to the fluid trajectory $(\bar{Q}(t), \bar{T}(t))$, when the stochastic network operates under the discipline $\mathcal{U}$.

From the fluid equation (16) we have $\sum_{(i,j)\in\sigma} \bar{T}_{i,j}(t_m, t_{m+1}) \le t_{m+1} - t_m$, for each m. As a result, any policy $\mathcal{U}$ is feasible. From the description above, it is certainly nonidling. We now analyze the dynamics of our network when policy $\mathcal{U}$ is implemented. For convenience, we introduce $s_{-1} \equiv s_0 = 0$.

LEMMA 8. *Under the policy $\mathcal{U}$ (in fact, under any scheduling policy), for every $m = 0, 1, \ldots, \lceil \frac{\theta}{\delta} \rceil$,*

$$\sup_{t_m \le t \le t_{m+1}} \|\bar{Q}(t) - \bar{Q}(t_m)\| \le C\delta n \tag{33}$$

and

$$\mathbb{P}\Big\{\sup_{t_m \le t \le t_{m+1}} \|\mathbf{Q}(t) - \mathbf{Q}(t_m)\| > C\delta n\Big\} \le O\big(\exp(-\Theta(n))\big). \tag{34}$$

PROOF. Applying (12), (13) and (16), we have

$$\|\bar{Q}(t) - \bar{Q}(t_m)\| \le \Big(\sum_i \lambda_i + \sum_i \mu_{i,J_i}\Big)(t - t_m) < C(t_{m+1} - t_m) = C\delta n,$$

which proves (33). We now prove (34). By Assumption B [specifically bound (10)], for every i and every $t \in [t_m, t_{m+1}]$,

$$\mathbb{P}\{|\mathbf{A}_i(t) - \mathbf{A}_i(t_m)| > 2\lambda_i\delta n\} \le \mathbb{P}\{|\mathbf{A}_i(t_{m+1}) - \mathbf{A}_i(t_m)| > 2\lambda_i\delta n\} \tag{35}$$

$$\le O\big(\exp(-\Theta(n))\big), \tag{36}$$

since $t_{m+1} - t_m = \delta n$. Similarly, for all i and j and $t \in [t_m, t_{m+1}]$,

$$\mathbb{P}\{|\mathbf{D}_{i,j}(t) - \mathbf{D}_{i,j}(t_m)| > 2\mu_{i,j}\delta n\} \le O\big(\exp(-\Theta(n))\big). \tag{37}$$

Applying (1) and (2), we obtain

$$\mathbb{P}\left\{\sup_{t_m \le t \le t_{m+1}} |\mathbf{Q}_{i,j}(t) - \mathbf{Q}_{i,j}(t_m)| > 2(\lambda_i + \mu_{i,j-1} + \mu_{i,j})\delta n\right\} \le O\big(\exp(-\Theta(n))\big).$$

By summing these probabilities over all (i, j), we obtain

$$\mathbb{P}\left\{\sup_{t_m \le t \le t_{m+1}} \|\mathbf{Q}(t) - \mathbf{Q}(t_m)\| > \left(2\sum_i \lambda_i J_i + 4\sum_{i,j} \mu_{i,j}\right)\delta n\right\}$$

$$\le \sum_{i,j} O\big(\exp(-\Theta(n))\big) = O\big(\exp(-\Theta(n))\big),$$

implying

$$\mathbb{P}\left\{\sup_{t_m \le t \le t_{m+1}} \|\mathbf{Q}(t) - \mathbf{Q}(t_m)\| > C\delta n\right\} \le O\big(\exp(-\Theta(n))\big),$$

which is (34). □

A large part of the remainder of the paper is devoted to proving Proposition 2 below. The proof is quite lengthy and we split the argument into several sections.

PROPOSITION 2. *Under the policy $\mathcal{U} = \mathcal{U}(\delta)$, for every $r = -1, 0, 1, \dots, M-1$, every $t_m \in [s_r, s_{r+1}]$ and every class (i, j),*

$$\mathbb{P}\{|\mathbf{Q}_{i,j}(t_m) - \bar{Q}_{i,j}(t_m)| \le \delta C^{r+3} n\} \ge 1 - O\big(\exp(-\Theta(n))\big). \tag{38}$$

The proof is done by using various induction steps. The "outer" induction is on r, which indexes the trajectory decomposition points s_r. The "inner" induction is done on the stages j of the classes (i, j) classes in the network, and is outlined in various lemmas below.

We start the outer induction with $r = -1$. Then for $t_m \in [s_{-1}, s_0] = \{0\}$, we simply have $t_m = 0$ and the bound in (38) holds trivially for all classes (i, j) since $\mathbf{Q}(0) = \bar{Q}(0) = q$, with probability one. Next we suppose the bounds in (38) hold for $-1, 0, 1, \dots, r-1$. We then show that the bounds hold for r. The necessary bounds will be established by a sequence of lemmas. Our first lemma simply says that assuming the bounds (38) hold for all $r' \le r-1$ and $t_m \in [s_{r'}, s_{r'+1}]$, a similar bound holds at the end point s_r.

LEMMA 9. *If the bound* (38) *holds for all $r' \le r-1$, then, for every i, j,*

$$\mathbb{P}\{|\mathbf{Q}_{i,j}(s_r) - \bar{Q}_{i,j}(s_r)| > \delta C^{r+2} n + 2\delta C n\} \le O\big(\exp(-\Theta(n))\big).$$

PROOF. Find the largest $t_{m'} \leq s_r$. Then $t_{m'+1} = t_{m'} + \delta n > s_r \geq t_{m'}$. By Lemma 8, we have $|\bar{Q}_{i,j}(s_r) - \bar{Q}_{i,j}(t_{m'})| \leq C\delta n$ and

$$\mathbb{P}\{|\mathbf{Q}_{i,j}(s_r) - \mathbf{Q}_{i,j}(t_{m'})| \geq C\delta n\} \leq O\bigl(\exp(-\Theta(n))\bigr).$$

Since $t_{m'} \in [s_{r-1}, s_r]$, then by the assumption of our induction in r,

$$\mathbb{P}\{|\mathbf{Q}_{i,j}(t_{m'}) - \bar{Q}_{i,j}(t_{m'})| \geq C^{r+2}\delta n\} \leq O\bigl(\exp(-\Theta(n))\bigr).$$

Combining the last three inequalities, we obtain the result. □

In Section 4.3.2 we obtain probabilistic lower bounds on the number of jobs processed during the time interval $[s_r, t_m)$, for any $t_m \in [s_r, s_{r+1}]$, under the scheduling policy $\mathcal{U}(\delta)$.

4.3.2. *Lower bounds on the departure process.* The next lemma shows that, with high probability, in the first stage in the route of each job type the total number of jobs processed during the time interval $[s_r, t_m]$ is not too far behind the amount of fluid processed during the same time interval in the fluid solution. A subsequent lemma establishes a similar bound for stages two and higher. Recall that we fixed r and we assume by induction that (38) holds for $r' \leq r - 1$.

LEMMA 10. *For every $i \leq I$ and every m such that $s_r \leq t_m \leq s_{r+1}$,*

$$\mathbb{P}\{\mathbf{D}_{i,1}(s_r, t_m) \geq \bar{D}_{i,1}(s_r, t_m) - 2\delta C^{r+2} n\} \geq 1 - O\bigl(\exp(-\Theta(n))\bigr) \tag{39}$$

and

$$\mathbb{P}\{\mathbf{T}_{i,1}(s_r, t_m) \geq \bar{T}_{i,1}(s_r, t_m) - 3\mu_{\max}\delta C^{r+2} n\} \geq 1 - O\bigl(\exp(-\Theta(n))\bigr). \tag{40}$$

PROOF. We start with proving bound (39). Bound (40) will be an easy corollary.

Part I. Fix a specific class $(i, 1)$, and time $t_{m_0}, s_r \leq t_{m_0} \leq s_{r+1}$, and introduce the event

$$\mathcal{D}(t_{m_0}) \equiv \bigl\{\mathbf{D}_{i,1}\bigl(s_r, t_{m_0}\bigr) < \bar{D}_{i,1}\bigl(s_r, t_{m_0}\bigr) - 2\delta C^{r+2} n\bigr\}. \tag{41}$$

Note then that (39) is equivalent to having $\mathbb{P}\{D(t_{m_0})\} \leq O(\exp(-\Theta(n)))$ for every i and $t_{m_0} \in [s_r, s_{r+1}]$. Next, we introduce the events

$$\mathcal{A} \equiv \{\forall\, t_m \in [s_r, s_{r+1}] : \mathbf{A}_i(s_r, t_m) \geq \lambda_i(t_m - s_r) - \delta C n\}, \tag{42}$$

$$\mathcal{Q} \equiv \{\mathbf{Q}_{i,1}(s_r) \geq \bar{Q}_{i,1}(s_r) - \delta C^{r+2} n - 2\delta C n\}. \tag{43}$$

From Lemma 9 and the inductive assumption, we have

$$\mathbb{P}\{\mathcal{Q}\} \geq 1 - O\bigl(\exp(-\Theta(n))\bigr). \tag{44}$$

Now fix any $t_m \in [s_r, s_{r+1}]$ and consider

$$\begin{aligned}&\mathbb{P}\{\mathbf{A}_i(s_r, t_m) \ge \lambda_i(t_m - s_r) - \delta C n\}\\ &\quad = \mathbb{P}\left\{\mathbf{A}_i(s_r, t_m) \ge \lambda_i(t_m - s_r) - \frac{\delta C n}{t_m - s_r}(t_m - s_r)\right\},\end{aligned}$$

where without loss of generality we may assume $t_m > s_r$. If $t_m - s_r \le \delta n$, then the probability above is equal to one, since the right-hand side of the inequality inside the probability is negative. Suppose now $t_m - s_r \ge \delta n$. We have

$$\frac{\delta C n}{t_m - s_r} \ge \frac{\delta C n}{\theta n} = \frac{\delta C}{\theta}.$$

Setting $\varepsilon = \frac{\delta C}{\theta}$ and using the large deviations Assumption B with this ε, we obtain that

$$\begin{aligned}\mathbb{P}\{\mathbf{A}_i(s_r, t_m) \ge \lambda_i(t_m - s_r) - \delta C n\} &\ge 1 - O\bigl(\exp(-\Theta(t_m - s_r))\bigr)\\ &\ge 1 - O\bigl(\exp(-\Theta(n))\bigr),\end{aligned} \tag{45}$$

where $t_m - s_r \ge \delta n$ is used in the last inequality. The number of different t_m in $[s_r, s_{r+1}]$ is at most $\theta n/(\delta n) = \theta/\delta$. Summing over all such t_m, we conclude

$$\mathbb{P}\{\mathcal{A}\} \ge 1 - (\theta/\delta) O\bigl(\exp(-\Theta(n))\bigr) = 1 - O\bigl(\exp(-\Theta(n))\bigr). \tag{46}$$

Hence,

$$\begin{aligned}\mathbb{P}\{\mathcal{D}(t_{m_0})\} &= \mathbb{P}\{\mathcal{D}(t_{m_0}) | \mathcal{A} \cap \mathcal{Q}\}\mathbb{P}\{\mathcal{A} \cap \mathcal{Q}\} + \mathbb{P}\{\mathcal{D}(t_{m_0}) | \overline{\mathcal{A} \cap \mathcal{Q}}\}\mathbb{P}\{\overline{\mathcal{A} \cap \mathcal{Q}}\}\\ &\le \mathbb{P}\{D(t_{m_0}) \cap A \cap Q\} + O\bigl(\exp(-\Theta(n))\bigr),\end{aligned}$$

where in the inequality we use $\mathbb{P}\{\overline{\mathcal{A} \cap \mathcal{Q}}\} \le O(\exp(-\Theta(n)))$, which holds by (44) and (46). Thus, to show (39), it suffices to prove

$$\mathbb{P}\{\mathcal{D}(t_{m_0}) \cap \mathcal{A} \cap \mathcal{Q}\} \le O\bigl(\exp(-\Theta(n))\bigr). \tag{47}$$

We denote the event $(\mathcal{D}(t_{m_0}) \cap \mathcal{A} \cap \mathcal{Q})$ by $\mathcal{D}_c(t_{m_0})$. We first show that given $\mathcal{D}_c(t_{m_0})$, there exists, with probability one, a time instance t_m with $s_r \le t_m \le t_{m_0}$, such that the following events occur:

$$\mathcal{F}(t_m) \equiv \{\mathbf{Q}_{i,1}(t_m) \ge \delta C n\} \tag{48}$$

and

$$\mathcal{G}(t_m) = \left\{\mathbf{D}_{i,1}(t_m, t_{m+1}) \le \mu_{i,1}\bar{T}_{i,1}(t_m, t_{m+1}) - \frac{2\delta^2 C n}{\theta}\right\}. \tag{49}$$

That is, we claim

$$\mathbb{P}\left\{\bigcup_{\{m\,:\, s_r \le t_m \le t_{m_0}\}} (\mathcal{F}(t_m) \cap \mathcal{G}(t_m)) \mid \mathcal{D}_c(t_{m_0})\right\} = 1. \tag{50}$$

On the other hand, we also claim that, for each $t_m \in [s_r, s_{r+1}]$,

$$\mathbb{P}\{\mathcal{G}(t_m) \mid \mathcal{F}(t_m)\} \le O\bigl(\exp(-\Theta(n))\bigr), \tag{51}$$

which implies that $\mathbb{P}(\mathcal{F}(t_m) \cap \mathcal{G}(t_m)) \le O(\exp(-\Theta(n)))$. Together with (50), this would imply

$$\begin{aligned}\mathbb{P}\{\mathcal{D}_c(t_{m_0})\} &\le \frac{\mathbb{P}\{\bigcup_{s_r \le t_m \le t_{m_0}} \mathcal{F}(t_m) \cap \mathcal{G}(t_m)\}}{\mathbb{P}\{\bigcup_{s_r \le t_m \le t_{m_0}} \mathcal{F}(t_m) \cap \mathcal{G}(t_m) \mid \mathcal{D}_c(t_{m_0})\}} \\ &\le \sum_{t_m \in [s_r, s_{r+1}]} \frac{\mathbb{P}\{\mathcal{F}(t_m) \cap \mathcal{G}(t_m)\}}{\mathbb{P}\{\bigcup_{s_r \le t_m \le t_{m_0}} \mathcal{F}(t_m) \cap \mathcal{G}(t_m) \mid \mathcal{D}_c(t_{m_0})\}} \\ &\le O\bigl(\exp(-\Theta(n))\bigr),\end{aligned}$$

where again, for the last inequality, we use the fact that the number of t_m in the interval $[s_r, s_{r+1}]$ is at most θ/δ and $\theta/\delta \exp(-\Theta(n)) = \exp(-\Theta(n))$. We could then conclude that (47) holds and we would be done. Thus, we need to show (50) and (51). We start by proving (51). Note that during the time interval I_m, policy $\mathcal{U}(\delta)$ either allocates at least $\bar{T}_{i,1}(t_m, t_{m+1})$ time units to process class $(i, 1)$ jobs, or all the $\mathbf{Q}_{i,1}(t_m) > \delta Cn$ jobs initially present are processed. In the second case $\mathcal{G}(t_m)$ does not hold since $\delta Cn > \mu_{i,1}\bar{T}_{i,1}(t_m, t_{m+1})$. In the first case, if $\bar{T}_{i,1}(t_m, t_{m+1}) < \frac{2\delta^2 Cn}{\mu_{i,1}\theta}$, then $\mathcal{G}(t_m)$ obviously does not hold, since the right-hand side in the inequality in (49) is negative. Otherwise, $\bar{T}_{i,1}(t_m, t_{m+1}) \ge \Theta(n)$. In this case, we can apply the large deviations bound (10) which holds by Assumption B. Setting $\varepsilon = 2\delta^2 Cn/(\bar{T}_{i,1}(t_m, t_{m+1})\theta) \ge 2\delta C/\theta$ in the bound, we obtain

$$\begin{aligned}\mathbb{P}\biggl\{\mathbf{D}_{i,1}(t_m, t_{m+1}) < \mu_{i,1}\bar{T}_{i,1}(t_m, t_{m+1}) - \frac{2\delta^2 Cn}{\theta}\biggr\} &\le O\bigl(\exp\bigl(-\Theta(\bar{T}_{i,1}(t_m, t_{m+1}))\bigr)\bigr) \\ &= O\bigl(\exp(-\Theta(n))\bigr),\end{aligned}$$

where in the last equation we use $\bar{T}_{i,1}(t_m, t_{m+1}) \ge \Theta(n)$ and as usual, δ, C and θ are hidden in the $\Theta(\cdot)$ notation. We conclude that (51) holds.

We now prove (50). Note that if $t_{m_0} - s_r < \delta n$, then the right-hand side of the inequality in the event $\mathcal{D}(t_{m_0})$ is negative and, therefore, the events $\mathcal{D}(t_{m_0})$ and $\mathcal{D}_c(t_{m_0})$ cannot occur. Thus, we assume there exists at least one $t_m \in [s_r, t_{m_0})$. We have $\mathbf{D}_{i,1}(s_r, t_{m_0}) \ge \sum_{\{m : s_r \le t_m \le t_{m_0 - 1}\}} \mathbf{D}_{i,1}(t_m, t_{m+1})$ and

$$\begin{aligned}\bar{D}_{i,1}(s_r, t_{m_0}) &\le \sum_{\{m : s_r - \delta n \le t_m \le t_{m_0 - 1}\}} \bar{D}_{i,1}(t_m, t_{m+1}) \\ &\le \sum_{\{m : s_r \le t_m \le t_{m_0 - 1}\}} \bar{D}_{i,1}(t_m, t_{m+1}) + \delta Cn.\end{aligned}$$

The event $\mathcal{D}(t_{m_0})$ implies that there exists a $t_m \in [s_r, t_{m_0-1}]$ such that

$$\begin{aligned}
\mathbf{D}_{i,1}(t_m, t_{m+1}) &< \mu_{i,1}\bar{T}_{i,1}(t_m, t_{m+1}) - \frac{2\delta C^{r+2}n - \delta C n}{\lceil (t_{m_0} - s_r)/\delta n \rceil} \\
&\le \mu_{i,1}\bar{T}_{i,1}(t_m, t_{m+1}) - \frac{2\delta C^{r+2}n - \delta C n}{(\theta/\delta) + 1} \qquad (52)\\
&\le \mu_{i,1}\bar{T}_{i,1}(t_m, t_{m+1}) - \frac{2\delta^2 C^{r+2}n - \delta^2 C}{\theta + \delta}, \qquad (53)\\
&\le \mu_{i,1}\bar{T}_{i,1}(t_m, t_{m+1}) - \frac{2\delta^2 C n}{\theta},
\end{aligned}$$

where we used $t_{m_0} - s_r \le \theta n$ in (52) and we use in (53) the fact that $\theta \ge 1$ by (25), $\delta \le 1$ by (32) and, as a result, $\theta + \delta \le 2\theta$ and $(2C^{r+2} - C)/2 \ge (2C^2 - C)/2 > 2C$.

Among $t_m \in [s_r, t_{m_0})$, select the largest m such that $\mathbf{D}_{i,1}(t_m, t_{m+1}) \le \mu_{i,1}\bar{T}_{i,1} \times (t_m, t_{m+1}) - \frac{2\delta^2 C n}{\theta}$ and denote it by $\hat{m}$. By the derivation above, the set of such t_m is nonempty. Thus,

$$\mathbf{D}_{i,1}(t_{\hat{m}}, t_{\hat{m}+1}) \le \mu_{i,1}\bar{T}_{i,1}(t_{\hat{m}}, t_{\hat{m}+1}) - \frac{2\delta^2 C n}{\theta}. \qquad (54)$$

Moreover, if $\hat{m} < m_0 - 1$, then for all $\hat{m} < m \le m_0 - 1$, we have

$$\mathbf{D}_{i,1}(t_m, t_{m+1}) \ge \mu_{i,1}\bar{T}_{i,1}(t_m, t_{m+1}) - \frac{2\delta^2 C n}{\theta},$$

or

$$\begin{aligned}
\mathbf{D}_{i,1}(t_{\hat{m}+1}, t_{m_0}) &\ge \mu_{i,1}\bar{T}_{i,1}(t_{\hat{m}+1}, t_{m_0}) - \frac{(m_0 - \hat{m})2\delta^2 C n}{\theta} \\
&\ge \mu_{i,1}\bar{T}_{i,1}(t_{\hat{m}+1}, t_{m_0}) - 2\delta C n,
\end{aligned} \qquad (55)$$

where $m_0 - \hat{m} \le \theta/\delta$ is used. Note, that the bound (55) holds trivially if $\hat{m} = m_0 - 1$. Next, note that the event $\mathcal{D}(t_{m_0})$ jointly with (55) implies

$$\begin{aligned}
\mathbf{D}_{i,1}(s_r, t_{\hat{m}+1}) &= \mathbf{D}_{i,1}(s_r, t_{m_0}) - \mathbf{D}_{i,1}(t_{\hat{m}+1}, t_{m_0}) \\
&\le \mu_{i,1}\bar{T}_{i,1}(s_r, t_{m_0}) - 2\delta C^{r+2}n - \bigl(\mu_{i,1}\bar{T}_{i,1}(t_{\hat{m}+1}, t_{m_0}) - 2\delta C n\bigr) \\
&= \mu_{i,1}\bar{T}_{i,1}(s_r, t_{\hat{m}+1}) - 2\delta C^{r+2}n + 2\delta C n.
\end{aligned} \qquad (56)$$

Thus,

$$\begin{aligned}
\mathbf{D}_{i,1}(s_r, t_{\hat{m}}) &\le \mathbf{D}_{i,1}(s_r, t_{\hat{m}+1}) \\
&\le \mu_{i,1}\bar{T}_{i,1}(s_r, t_{\hat{m}+1}) - 2\delta C^{r+2}n + 2\delta C n \\
&\le \mu_{i,1}\bar{T}_{i,1}(s_r, t_{\hat{m}}) - 2\delta C^{r+2}n + 3\delta C n,
\end{aligned} \qquad (57)$$

where $\mu_{i,1}\bar{T}_{i,1}(t_{\hat{m}}, t_{\hat{m}+1}) \le \mu_{i,1}\delta n < \delta C n$ is used. Now recall from (1) that

$$\mathbf{Q}_{i,1}(t_{\hat{m}}) = \mathbf{Q}_{i,1}(s_r) + \mathbf{A}_i(s_r, t_{\hat{m}}) - \mathbf{D}_{i,1}(s_r, t_{\hat{m}}). \tag{58}$$

Then conditioned on $\mathcal{D}_c(t_{m_0}) = \mathcal{D}(t_{m_0}) \cap \mathcal{A} \cap \mathcal{Q}$ and using (57), we obtain

$$\begin{aligned}\mathbf{Q}_{i,1}(t_{\hat{m}}) \ge{}& \bar{Q}_{i,1}(s_r) + \lambda_i(t_{\hat{m}} - s_r) - \mu_{i,1}\bar{T}_{i,1}(s_r, t_{\hat{m}}) \\ &- \delta C^{r+2} n - 2\delta C n - \delta C n + 2\delta C^{r+2} n - 3\delta C n.\end{aligned}$$

Recall from (12) that $\bar{Q}_{i,1}(s_r) + \lambda_i(t_{\hat{m}} - s_r) - \mu_{i,1}\bar{T}_{i,1}(s_r, t_{\hat{m}}) = \bar{Q}_{i,1}(t_{\hat{m}}) \ge 0$. Then

$$\mathbf{Q}_{i,1}(t_{\hat{m}}) \ge \bar{Q}_{i,1}(t_{\hat{m}}) + \delta C^{r+2} n - 6\delta C n > \delta C n. \tag{59}$$

We have established that if the event $\mathcal{D}_c(t_{m_0})$ holds, then (54) and (59) hold for some $t_{\hat{m}} \le t_{m_0}$. In other words, (50) holds. This completes the proof of (39).

Part II. We now prove (40). Fix a $t_m \in [s_r, s_{r+1}]$. Note that the bound (40) is trivial if $\bar{T}_{i,1}(s_r, t_m) < 3\mu_{\max}\delta C^{r+2} n$. So, suppose the previous inequality does not hold. Let

$$\Delta = \mu_{i,1}\bar{T}_{i,1}(s_r, t_m) - 2\delta C^{r+2} n \ge \delta C^{r+2} n \ge \Theta(n), \tag{60}$$

where we use $\mu_{i,1}\mu_{\max} \ge 1$. Let

$$\varepsilon = \frac{\delta C n}{\Delta} \ge \frac{\delta}{\theta},$$

where we use $\Delta < \mu_{i,1}\bar{T}_{i,1}(s_r, t_m) < C\theta n$. We condition on the event $\mathcal{D}(t_m)$, which by (39) holds with probability at least $1 - \exp(-\Theta(n))$, and use large deviations Assumption B with the ε above to obtain

$$\begin{aligned}&\mathbb{P}\{\mathbf{T}_{i,1}(s_r, t_m) \ge \mu_{i,1}^{-1}(\mu_{i,1}\bar{T}_{i,1}(s_r, t_m) - 2\delta C^{r+2} n) - \delta C n \mid \mathcal{D}(t_m)\} \\ &\quad = \mathbb{P}\{\mathbf{T}_{i,1}(s_r, t_m) \ge \mu_{i,1}^{-1}\Delta - \varepsilon\Delta | \mathbf{D}(s_r, t_m) \ge \Delta\} \\ &\quad \ge 1 - O(\exp(-\Theta(\Delta))) \\ &\quad \ge 1 - O(\exp(-\Theta(n))),\end{aligned}$$

where the last inequality follows from the last inequality in (60). To finish the argument, we observe that $\mu_{i,1}^{-1} 2\delta C^{r+2} + \delta C n < 3\mu_{\max}\delta C^{r+2} n$. □

We now establish a similar lower bound for classes corresponding to stages two and higher.

LEMMA 11. *For every $i \le I$, $j \le J_i$ and m such that $s_r \le t_m \le s_{r+1}$,*

$$\mathbb{P}\{\mathbf{D}_{i,j}(s_r, t_m) \ge \bar{D}_{i,j}(s_r, t_m) - 2\delta j C^{r+2} n\} \ge 1 - \exp(-\Theta(n)) \tag{61}$$

and

$$\mathbb{P}\{\mathbf{T}_{i,j}(s_r, t_m) \ge \bar{T}_{i,j}(s_r, t_m) - 3\mu_{\max}\delta j C^{r+2} n\} \ge 1 - \exp(-\Theta(n)). \tag{62}$$

PROOF. The proof is very similar to the one for Lemma 10. We only highlight the differences. The proof is done by induction in j; the base case $j = 1$ is covered by Lemma 10. So let us fix a $j > 1$ and assume that the assertion holds for all (i, j') with $j' \le j - 1$. We again define an event related to the inequality inside $\mathbb{P}$ in (61). For a class (i, j) and any time t_{m_0} with $s_r \le t_{m_0} \le s_{r+1}$, let

$$\mathcal{D}(t_{m_0}) = \{\mathbf{D}_{i,j}(s_r, t_m) < \bar{D}_{i,j}(s_r, t_m) - 2\delta j C^{r+2} n\}. \tag{63}$$

We need to show $\mathbb{P}\{\mathcal{D}(t_{m_0})\} \le O(\exp(-\Theta(n)))$ for every $t_{m_0} \in [s_r, s_{r+1}]$. As in Lemma 10, we introduce the event

$$\mathcal{Q} \equiv \{\mathbf{Q}_{i,j}(s_r) \ge \bar{Q}_{i,j}(s_r) - \delta C^{r+2} n - 2\delta C n\}, \tag{64}$$

but instead of the event $\mathcal{A}$ defined by (42), consider

$$\mathcal{D} \equiv \{\forall t_m \in [s_r, s_{r+1}] : \mathbf{D}_{i,j-1}(s_r, t_m) \ge \bar{D}_{i,j-1}(s_r, t_m) - 2\delta(j-1) C^{r+2} n\}. \tag{65}$$

Again using Lemma 9 (and the "outer" inductive assumption), we obtain $\mathbb{P}\{\mathcal{Q}\} \ge 1 - \exp(-\Theta(n))$ and by the inductive assumption on j, $\mathbb{P}\{\mathcal{D}\} \ge 1 - O(\exp(-\Theta(n)))$ [where, as before, we sum several expressions of the order $O(\exp(-\Theta(n)))$ over $t_m \in [s_r, s_{r+1}]$ to get again $O(\exp(-\Theta(n)))$]. Next, let $\mathcal{D}_c(t_{m_0}) = \mathcal{D}(t_{m_0}) \cap \mathcal{Q} \cap \mathcal{D}$. We need to show $\mathbb{P}\{\mathcal{D}_c(t_{m_0})\} \le O(\exp(-\Theta(n)))$. For every $t_m \in [s_r, t_{m_0}]$, we introduce the event $\mathcal{F}(t_m)$ as in (48), except $\mathbf{Q}_{i,j}$ is used instead of $\mathbf{Q}_{i,1}$. Finally, we introduce $\mathcal{G}(t_m)$, defined as follows:

$$\mathcal{G}(t_m) = \left\{\mathbf{D}_{i,j}(t_m, t_{m+1}) \le \mu_{i,j} \bar{T}_{i,j}(t_m, t_{m+1}) - \frac{2\delta^2 C n j}{\theta}\right\}.$$

Arguing as in the proof of Lemma 10, we claim that (50) and (51) hold with the new event definitions. The proof of (51) is identical to the one of Lemma 10. For (50), we repeat the argument until we get to (57), instead of which we get

$$\mathbf{D}_{i,j}(s_r, t_{\hat{m}}) \le \mu_{i,1} \bar{T}_{i,j}(s_r, t_{\hat{m}}) - 2j\delta C^{r+2} n + 3\delta C n. \tag{66}$$

Then we obtain

$$\begin{align} \mathbf{Q}_{i,j}(t_{\hat{m}}) &= \mathbf{Q}_{i,j}(s_r) + \mathbf{D}_{i,j-1}(s_r, t_{\hat{m}}) - \mathbf{D}_{i,j}(s_r, t_{\hat{m}}) \tag{67}\\ &\ge \bar{Q}_{i,j}(s_r) + \mu_{i,j-1} \bar{T}_{i,j-1}(s_r, t_{\hat{m}}) - \mu_{i,j} \bar{T}_{i,j}(s_r, t_{\hat{m}}) \notag\\ &\quad - \delta C^{r+2} n - 2\delta C n - 2\delta(j-1) C^{r+2} n + 2\delta j C^{r+2} n - 3\delta C n \tag{68}\\ &= \bar{Q}_{i,j}(t_{\hat{m}}) + \delta C^{r+2} n - 5\delta C n \tag{69}\\ &\ge \delta C n, \notag \end{align}$$

where (2) is used for (67), conditioning on $\mathcal{D}_c(t_{m_0})$ is used in (68), and (13) is used in (69). This proves (50) and completes the proof of (61). The proof of the lower bound for $\mathbf{T}_{i,j}(\cdot)$ follows the proof of Lemma 10, almost line for line. □

4.3.3. *Upper bounds on the departure processes.* In this section we obtain upper bounds, similar to the bounds in Lemmas 10 and 11, on the cumulative departures $\mathbf{D}_{i,j}(s_r, t_m)$, for values of m such that $s_r < t_m \leq s_{r+1}$.

For every station σ, by construction of the sequence $s_0, s_1, \dots, s_M$, we have either $\bar{Q}_\sigma(t) > 0$ for all $s_r < t < s_{r+1}$ or $\bar{Q}_\sigma(t) = 0$ for all $s_r < t < s_{r+1}$. We consider these cases separately.

LEMMA 12. *Given any station σ, suppose the interval (s_r, s_{r+1}) is such that $\bar{Q}_\sigma(t) > 0$ for all $s_r < t < s_{r+1}$. Then, for every class $(i, j) \in \sigma$ and every $t_m \in [s_r, s_{r+1}]$,*

$$\begin{aligned}\mathbb{P}\{\mathbf{T}_{i,j}(s_r, t_m) \leq \bar{T}_{i,j}(s_r, t_m) + 3\delta|\sigma|\mu_{\max}J_{\max}C^{r+2}n\}\\ \geq 1 - O\bigl(\exp(-\Theta(n))\bigr)\end{aligned} \tag{70}$$

and

$$\begin{aligned}\mathbb{P}\{\mathbf{D}_{i,j}(s_r, t_m) \leq \bar{D}_{i,j}(s_r, t_m) + 4\delta|\sigma|\mu_{\max}^2J_{\max}C^{r+2}n\}\\ \geq 1 - O\bigl(\exp(-\Theta(n))\bigr).\end{aligned} \tag{71}$$

PROOF. Given any station σ, suppose that $\bar{Q}_\sigma(t) > 0$ for all $s_r < t < s_{r+1}$. By the nonidling constraint (20), we have that

$$\sum_{(i,j)\in\sigma} \bar{T}_{i,j}(s_r, s_{r+1}) = s_{r+1} - s_r. \tag{72}$$

Fix any $t_m \in [s_r, s_{r+1}]$ and fix any class $(i, j) \in \sigma$. Applying Lemmas 10 and 11 to s_r and t_m, we have that with probability at least $1 - O(\exp(-\Theta(n)))$, for every class $(i', j') \in \sigma$,

$$\begin{aligned}\mathbf{T}_{i',j'}(s_r, t_m) &\geq \bar{T}_{i',j'}(s_r, t_m) - 3\delta j'\mu_{\max}C^{r+2}n\\ &\geq \bar{T}_{i',j'}(s_r, t_m) - 3\delta\mu_{\max}J_{\max}C^{r+2}n,\end{aligned}$$

where we use (7) in the second inequality. Applying (72) and the feasibility inequality (4), we obtain that with probability at least $1 - O(\exp(-\Theta(n)))$,

$$\begin{aligned}\mathbf{T}_{i,j}(s_r, t_m) &\leq t_m - s_r - \sum_{(i,j)\neq(i',j')\in\sigma} \mathbf{T}_{i',j'}(s_r, t_m)\\ &\leq t_m - s_r - \sum_{(i,j)\neq(i',j')\in\sigma} \bigl(\bar{T}_{i',j'}(s_r, t_m) - 3\delta\mu_{\max}J_{\max}C^{r+2}n\bigr)\\ &\leq \bar{T}_{i,j}(s_r, t_m) + 3\delta|\sigma|\mu_{\max}J_{\max}C^{r+2}n.\end{aligned} \tag{73}$$

Let us define the event $\mathcal{T}$ as follows:

$$\mathcal{T} = \{\mathbf{T}_{i,j}(s_r, t_m) \leq \bar{T}_{i,j}(s_r, t_m) + 3\delta|\sigma|\mu_{\max}J_{\max}C^{r+2}n\},$$

that is, $\mathbb{P}\{\mathcal{T}\} \geq 1 - O(\exp(-\Theta(n)))$ per (73). Next, consider

$$\begin{aligned}\mathbb{P}\{&\mathbf{D}_{i,j}(s_r, t_m) > \mu_{i,j}\bar{T}_{i,j}(s_r, t_m) + 4\delta|\sigma|\mu_{\max}^2 J_{\max} C^{r+2} n \mid \mathcal{T}\}\\ &\leq \mathbb{P}\bigl\{\mathbf{D}_{i,j}(s_r, t_m) > \mu_{i,j}\bigl(\bar{T}_{i,j}(s_r, t_m) + 3\delta|\sigma|\mu_{\max} J_{\max} C^{r+2} n\bigr)\\ &\qquad\qquad + \delta|\sigma|\mu_{\max}^2 J_{\max} C^{r+2} n \mid \mathcal{T}\bigr\}.\end{aligned}$$

Applying the large deviation bound (10) with

$$\varepsilon = \frac{\delta|\sigma|\mu_{\max}^2 J_{\max} C^{r+2} n}{\bar{T}_{i,j}(s_r, t_m) + 3\delta|\sigma|\mu_{\max} J_{\max} C^{r+2} n} \geq \frac{\delta|\sigma|\mu_{\max}^2 J_{\max} C^{r+2}}{\theta + 3\delta|\sigma|\mu_{\max} J_{\max} C^{r+2}}$$

[where $\bar{T}_{i,j}(s_r, t_m) \leq \theta n$ is used], we obtain

$$\begin{aligned}\mathbb{P}\{&\mathbf{D}_{i,j}(s_r, t_m) > \mu_{i,j}\bar{T}_{i,j}(s_r, t_m) + 4\delta|\sigma|\mu_{\max}^2 J_{\max} C^{r+2} n \mid T\}\\ &\leq \exp\bigl(-\Theta\bigl(\bar{T}_{i,j}(s_r, t_m) + 3\delta|\sigma|\mu_{\max} J_{\max} C^{r+2} n\bigr)\bigr) \leq \exp(-\Theta(n)),\end{aligned}$$

where the constant ε is hidden in $\Theta(\cdot)$. Taking this together with $\mathbb{P}\{\mathcal{T}\} \geq 1 - O(\exp(-\Theta(n)))$, we have proven the lemma. □

We now analyze stations σ for which the fluid amount stays zero during the interval $[s_r, s_{r+1}]$. In the following lemma we obtain an analogue of Lemma 12 for this second case.

LEMMA 13. *Given any station σ, suppose the interval (s_r, s_{r+1}) is such that $\bar{Q}_\sigma(t) = 0$, for all $s_r < t < s_{r+1}$. Then for every $(i, j) \in \sigma$ and every $t_m \in [s_r, s_{r+1}]$,*

$$\begin{aligned}\mathbb{P}\{&\mathbf{D}_{i,j}(s_r, t_m) \leq \bar{D}_{i,j}(s_r, t_m) + 5\delta j \mu_{\max}^2 I J_{\max}^2 C^{r+2} n\}\\ &\geq 1 - O\bigl(\exp(-\Theta(n))\bigr).\end{aligned} \tag{74}$$

PROOF. Consider any station σ such that $\bar{Q}_\sigma(t) = 0$, for all $s_r < t < s_{r+1}$. Applying fluid equations (12) and (13), we obtain that, for every class $(i, j) \in \sigma$ and every $t_m \in [s_r, s_{r+1}]$,

$$\bar{D}_{i,j-1}(s_r, t_m) = \bar{D}_{i,j}(s_r, t_m), \tag{75}$$

where for the case $j = 1$, $\bar{D}_{i,j-1}(\cdot)$ is understood as $\bar{A}_i(\cdot)$.

The proof now proceeds by induction in j. We start with the base step, $j = 1$. So, consider any class $(i, 1) \in \sigma$. Applying (75), we have

$$\bar{D}_{i,1}(s_r, t_m) = \bar{A}_i(s_r, t_m). \tag{76}$$

Applying (1), we have

$$\mathbf{Q}_{i,1}(t_m) = \mathbf{Q}_{i,1}(s_r) + \mathbf{A}_i(s_r, t_m) - \mathbf{D}_{i,1}(s_r, t_m) \geq 0. \tag{77}$$

Next, let $t_{m'} \equiv \max\{m : t_m \le s_r\}$. In particular, $0 \le s_r - t_{m'} \le \delta n$. We then have

$$\begin{aligned} &\mathbb{P}\{\mathbf{A}_i(s_r, t_m) > \bar{A}_i(s_r, t_m) + 2\delta C n\} \\ &\qquad \le \mathbb{P}\{\mathbf{A}_i(t_{m'}, t_m) > \bar{A}_i(s_r, t_m) + 2\delta C n\} \end{aligned} \tag{78}$$

$$\qquad \le \mathbb{P}\{\mathbf{A}_i(t_{m'}, t_m) > \bar{A}_i(t_{m'} + \delta n, t_m) + 2\delta C n\} \tag{79}$$

$$\qquad \le \mathbb{P}\{\mathbf{A}_i(t_{m'}, t_m) > \bar{A}_i(t_{m'}, t_m) + \delta C n\}, \tag{80}$$

where $\bar{A}_i(t_{m'}, t_{m'} + \delta n) \le \lambda_i \delta n < C\delta n$ is used in (80). Note that when $t_{m'} = t_m$, the probability in (80) is zero since the left-hand side is negative. Thus, we assume $t_{m'} > t_m$. Let $\varepsilon = \delta C n/(t_m - t_{m'}) \ge \delta C n/(\theta n) = \delta C/\theta > 0$. Using the large deviations Assumption B with this ε, we obtain

$$\begin{aligned} \mathbb{P}\{\mathbf{A}_i(t_{m'}, t_m) > \bar{A}_i(t_{m'}, t_m) + \delta C n\} &= \mathbb{P}\{\mathbf{A}_i(t_{m'}, t_m) > \bar{A}_i(t_{m'}, t_m) + \varepsilon(t_m - t_{m'})\} \\ &\le O\bigl(\exp(-\Theta(t_m - t_{m'}))\bigr) \\ &\le O\bigl(\exp(-\Theta(n))\bigr), \end{aligned}$$

where $t_m - t_{m'} \ge \delta n$ is used in the last inequality. Combining this bound with (80), we obtain

$$\mathbb{P}\{\mathbf{A}_i(s_r, t_m) > \bar{A}_i(s_r, t_m) + 2\delta C n\} \le O\bigl(\exp(-\Theta(n))\bigr). \tag{81}$$

Applying Lemma 9, we have

$$\mathbb{P}\{\mathbf{Q}_{i,1}(s_r) \le \bar{Q}_{i,1}(s_r) + \delta C^{r+2} n + 2\delta C n\} \ge 1 - O\bigl(\exp(-\Theta(n))\bigr).$$

By our assumption that $\bar{Q}_\sigma(t) = 0$ for all $s_r < t < s_{r+1}$ and by continuity, we have $\bar{Q}_\sigma(s_r) = 0$. Using this fact, we now have

$$\mathbb{P}\{\mathbf{Q}_{i,1}(s_r) \le \delta C^{r+2} n + 2\delta C n\} \ge 1 - O\bigl(\exp(-\Theta(n))\bigr). \tag{82}$$

Now, from (77), we have

$$\begin{aligned} \mathbf{D}_{i,1}(s_r, t_m) &= \mathbf{Q}_{i,1}(s_r) + \mathbf{A}_i(s_r, t_m) - \mathbf{Q}_{i,1}(t_m) \\ &\le \mathbf{Q}_{i,1}(s_r) + \mathbf{A}_i(s_r, t_m). \end{aligned}$$

Applying (81) and (82), we obtain that, with probability at least $1 - O(\exp(-\Theta(n)))$,

$$\begin{aligned} \mathbf{D}_{i,1}(s_r, t_m) &\le \bar{A}_i(s_r, t_m) + 2\delta C n + \delta C^{r+2} n + 2\delta C n \\ &\le \bar{A}_i(s_r, t_m) + 5\delta \mu_{\max}^2 I J_{\max}^2 C^{r+2} n. \end{aligned}$$

Combining this with (76), we obtain the required bound. This completes the proof of the base step.

We now prove the inductive step. So, fix $j > 1$ and suppose that the assertion holds for $1, 2, \dots, j-1$. We now consider a particular class $(i, j) \in \sigma$. We have, from (2),

$$\mathbf{Q}_{i,j}(t_m) = \mathbf{Q}_{i,j}(s_r) + \mathbf{D}_{i,j-1}(s_r, t_m) - \mathbf{D}_{i,j}(s_r, t_m) \ge 0. \tag{83}$$

Again, by Lemma 9, we have

$$\mathbb{P}\{\mathbf{Q}_{i,j}(s_r) \le \bar{Q}_{i,j}(s_r) + \delta C^{r+2}n + 2\delta Cn\} \ge 1 - O\big(\exp(-\Theta(n))\big),$$

which implies

$$\mathbb{P}\{\mathbf{Q}_{i,j}(s_r) \le \delta C^{r+2}n + 2\delta Cn\} \ge 1 - O\big(\exp(-\Theta(n))\big), \tag{84}$$

again since $\bar{Q}_\sigma(s_r) = 0$.

Consider the station $\sigma_{\nu'}$ containing $(i, j-1)$. If $\sigma_{\nu'}$ is also such that $\bar{Q}_{\sigma'}(t) = 0$ for all $s_r < t < s_{r+1}$ (e.g., when $\nu' = \nu$), then by the inductive assumption on j,

$$\mathbb{P}\{\mathbf{D}_{i,j-1}(s_r, t_m) \le \bar{D}_{i,j-1}(s_r, t_m) + 5\delta(j-1)\mu^2_{\max} I J^2_{\max} C^{r+2}n\}$$
$$\ge 1 - O\big(\exp(-\Theta(n))\big).$$

Otherwise, $\sigma_{\nu'}$ is such that $\bar{Q}_{\sigma_{\nu'}}(t) > 0$ for all $s_r < t < s_{r+1}$. Then Lemma 12 becomes applicable, and applying (71) to $\sigma_{\nu'}$, we have that, with probability at least $1 - O(\exp(-\Theta(n)))$,

$$\begin{aligned}\mathbf{D}_{i,j-1}(s_r, t_m) &\le \bar{D}_{i,j-1}(s_r, t_m) + 4\delta|\sigma_{\nu'}|\mu^2_{\max} J_{\max} C^{r+2}n \\ &< \bar{D}_{i,j-1}(s_r, t_m) + 5(j-1)\delta\mu^2_{\max} I J^2_{\max} C^{r+2}n,\end{aligned}$$

where we use $|\sigma_{\nu'}| \le \sum_i J_i \le I J_{\max}$ and $j > 1$. Hence, in either case we have a probabilistic bound on $\mathbf{D}_{i,j-1}(s_r, t_m)$. Combining this with (83) and (84), we obtain that, with probability at least $1 - O(\exp(-\Theta(n)))$,

$$\begin{aligned}\mathbf{D}_{i,j}(s_r, t_m) &\le \bar{D}_{i,j-1}(s_r, t_m) + \delta C^{r+2}n + 2\delta Cn + 5\delta(j-1)\mu^2_{\max} I J^2_{\max} C^{r+2}n \\ &< \bar{D}_{i,j-1}(s_r, t_m) + \delta\big(3 + 5(j-1)\mu^2_{\max} I J^2_{\max}\big) C^{r+2}n \\ &< \bar{D}_{i,j-1}(s_r, t_m) + 5\delta j\mu^2_{\max} I J^2_{\max} C^{r+2}n.\end{aligned}$$

Finally, recalling (75), we obtain the desired bound. This completes the proof of the inductive step. □

With the lemmas above in hand, we are now ready to finish the proof of Proposition 2, by completing the outer inductive step on r.

PROOF OF PROPOSITION 2. Fix any $t_m \in [s_r, s_{r+1}]$. By Lemma 9,

$$\mathbb{P}\{|\mathbf{Q}_{i,j}(s_r) - \bar{Q}_{i,j}(s_r)| \le \delta C^{r+2}n + 2\delta Cn\} \ge 1 - O\big(\exp(-\Theta(n))\big).$$

Next, for any class (i, j), recall that we have

$$\mathbf{Q}_{i,j}(t_m) = \mathbf{Q}_{i,j}(s_r) + \mathbf{D}_{i,j-1}(s_r, t_m) - \mathbf{D}_{i,j}(s_r, t_m), \tag{85}$$

with $\mathbf{D}_{i,j-1}(\cdot)$ replaced by $\mathbf{A}_i(\cdot)$ when $j = 1$. Combining Lemmas 10, 11, 12 and 13, we obtain that

$$\mathbb{P}\{|\mathbf{D}_{i,j}(s_r, t_m) - \bar{D}_{i,j}(s_r, t_m)| \le 5j\delta\mu^2_{\max} I J^2_{\max} C^{r+2}n\}$$
$$\ge 1 - O\big(\exp(-\Theta(n))\big).$$

Recalling (45) and (81), we have that

$$\mathbb{P}\{|\mathbf{A}_i(s_r,t_m)-\bar{A}_i(s_r,t_m)|\le 2\delta Cn\}\ge 1-O\bigl(\exp(-\Theta(n))\bigr).$$

Combining the previous two bounds with (85) and the fluid analogs (12) and (13), we obtain that, with probability at least $1-O(\exp(-\Theta(n)))$,

$$|\mathbf{Q}_{i,j}(t_m)-\bar{Q}_{i,j}(t_m)|\le \delta C^{r+2}n+2\delta Cn+10\delta j\mu_{\max}^2 IJ_{\max}^2C^{r+2}n<\delta C^{r+3}n,$$

where the bound (8) is used. This completes the proof of Proposition 2. □

4.4. *Proof of the main theorem.* In this section we present the final two proofs. The next theorem is the last result needed before proving Theorem 7.

THEOREM 14. *Suppose the fluid model of a stochastic multitype network is not globally weakly stable and satisfies the FDP property. Then for any initial state* $\mathbf{Q}(0)=(q,z_1,z_2)\in\mathbb{Z}_+^d\times\Re_+^{I+d}$, *under the nonidling scheduling policy* $\mathcal{U}$, *we have*

$$\mathbb{P}\{\|\mathbf{Q}(\theta\|q\|)\|\ge 2\|q\|\}\ge 1-O\bigl(\exp(-\Theta(\|q\|))\bigr) \tag{86}$$

and

$$\mathbb{P}\left\{\inf_{0\le t\le\theta\|q\|}\|\mathbf{Q}(t)\|\ge\frac{\|q\|}{4}\max\left(\frac{\gamma}{C},1\right)\right\}\ge 1-O\bigl(\exp(-\Theta(\|q\|))\bigr). \tag{87}$$

PROOF. We first prove (87). Fix any t_m and find the s_r such that $t_m\in[s_r,s_{r+1}]$. We have

$$\begin{aligned}\mathbb{P}&\left\{\sup_{t_m\le t\le t_{m+1}}\|\mathbf{Q}(t)-\bar{Q}(t)\|>3C^{r+4}\delta n\right\}\\ &\le\mathbb{P}\left\{\sup_{t_m\le t\le t_{m+1}}\|\mathbf{Q}(t)-\mathbf{Q}(t_m)\|>C\delta n\right\}\\ &\quad+\mathbb{P}\{\|\mathbf{Q}(t_m)-\bar{Q}(t_m)\|>C^{r+4}\delta n\}\\ &\quad+\mathbb{P}\left\{\sup_{t_m\le t\le t_{m+1}}\|\bar{Q}(t_m)-\bar{Q}(t)\|>C\delta n\right\}\\ &\le O\bigl(\exp(-\Theta(n))\bigr).\end{aligned}$$

In fact, observe that the last probability in the right-hand side above is equal to zero by (33) of Lemma 8. The first probability in the right-hand side is at most $O(\exp(-\Theta(n)))$ by (34) of Lemma 8, and the second probability is also at most $O(\exp(-\Theta(n)))$ by (38) of Proposition 2 and the fact that $\sum J_i<C$.

Combining the inequality above with (31), we obtain

$$\mathbb{P}\left\{\inf_{t_m\le t\le t_{m+1}}\|\mathbf{Q}(t)\|<\frac{n}{2}\min\left(\frac{\gamma}{C},1\right)-3C^{r+4}\delta n\right\}\le O\bigl(\exp(-\Theta(n))\bigr).$$

From (32) and since $r \le M-1$, we have

$$\frac{n}{2}\min\left(\frac{\gamma}{C},1\right) - 3C^{r+4}\delta n \ge \frac{n}{4}\min\left(\frac{\gamma}{C},1\right).$$

Thus,

$$\mathbb{P}\left\{\inf_{t_m \le t \le t_{m+1}} \|\mathbf{Q}(t)\| < \frac{n}{4}\min\left(\frac{\gamma}{C},1\right)\right\} \le O\big(\exp(-\Theta(n))\big).$$

By summing over all $m = 0,1,\dots,\lceil\theta/\delta\rceil$, we obtain

$$\mathbb{P}\left\{\inf_{0\le t\le\theta n} \|\mathbf{Q}(t)\| < \frac{n}{4}\min\left(\frac{\gamma}{C},1\right)\right\} \le \left\lceil\frac{\theta}{\delta}\right\rceil O\big(\exp(-\Theta(n))\big) = O\big(\exp(-\Theta(n))\big),$$

where the last equality follows since by (25) and (32), the value of $\lceil\theta/\delta\rceil$ is bounded above by a constant. Recall, finally, that $\|q\| = n$. This completes the proof of (87).

We now prove (86). Find the largest $t_m \le \theta n$. In particular, $\theta n - t_m \le \delta n$. Applying (34) with $t = \theta n$, we obtain

$$\mathbb{P}\{\|\mathbf{Q}(\theta n) - \mathbf{Q}(t_m)\| > C\delta n\} \le O\big(\exp(-\Theta(n))\big). \tag{88}$$

Applying (33) at $t = \theta n$, we obtain

$$\|\bar{Q}(\theta n) - \bar{Q}(t_m)\| \le C\delta n. \tag{89}$$

Applying (38) to the t_m chosen above, we obtain

$$\mathbb{P}\{|\mathbf{Q}_{i,j}(t_m) - \bar{Q}_{i,j}(t_m)| > \delta C^{M+2} n\} \le O\big(\exp(-\Theta(n))\big). \tag{90}$$

Next, we note that

$$\begin{aligned}
&\mathbb{P}\{\|\mathbf{Q}(t_m) - \bar{Q}(t_m)\| > \delta C^{M+3} n\}\\
&\quad\le \mathbb{P}\left\{\bigcup_{i,j} |\mathbf{Q}_{i,j}(t_m) - \bar{Q}_{i,j}(t_m)| > \frac{\delta C^{M+3} n}{I J_{\max}}\right\}\\
&\quad\le \mathbb{P}\left\{\bigcup_{i,j} |\mathbf{Q}_{i,j}(t_m) - \bar{Q}_{i,j}(t_m)| > \frac{\delta C^{M+3} n}{C}\right\}\\
&\quad= \mathbb{P}\left\{\bigcup_{i,j} |\mathbf{Q}_{i,j}(t_m) - \bar{Q}_{i,j}(t_m)| > \delta C^{M+2} n\right\}\\
&\quad\le \sum_{i,j} \mathbb{P}\{|\mathbf{Q}_{i,j}(t_m) - \bar{Q}_{i,j}(t_m)| > \delta C^{M+2} n\}\\
&\quad\le O\big(\exp(-\Theta(n))\big).
\end{aligned}$$

In the last step, we employ (90) and then sum over all i and j to obtain a new exponential bound. Combining (88), (89) and the last bound, we obtain

$$\mathbb{P}\{\|\mathbf{Q}(\theta n) - \bar{Q}(\theta n)\| > 3\delta C^{M+3} n\} \le O\big(\exp(-\Theta(n))\big).$$

Since $\delta < 1/(3C^{M+3})$, we obtain

$$\mathbb{P}\{\|\mathbf{Q}(n\theta) - \bar{Q}(n\theta)\| > n\} \le O\bigl(\exp(-\Theta(n))\bigr).$$

Recalling from (30) that $\|\bar{Q}(n\theta)\| \ge 3n$ and recalling $\|q\| = n$, we obtain

$$\mathbb{P}\{\|\mathbf{Q}(n\theta)\| < 2\|q\|\} \le O\bigl(\exp(-\Theta(n))\bigr),$$

which implies (86). This completes the proof of Theorem 14. □

It should be noted that the constant "2" which appears in (86) is completely arbitrary. In all of the proofs in which the constant appears, it can be replaced by any constant greater than unity. We are now ready to prove our main result.

PROOF OF THEOREM 7. We fix a large value n_0 (the actual value will be specified later). Consider any initial state (q, z_1, z_2) with $\|q\| \ge n_0$. We apply the policy $\mathcal{U}$ for the time interval $[0, \theta_0]$, where $\theta_0 = \theta\|q\|$. If at time θ_0 the resulting state $\mathbf{Q}(\theta_0)$ is such that $\|\mathbf{Q}(\theta_0)\| \ge 2n_0$, then we apply the policy $\mathcal{U}$ again with q reset to $\mathbf{Q}(\theta_0)$, till the corresponding time $\theta_1 = \theta_0 + \theta\|\mathbf{Q}(\theta_0)\|$. If again $\|\mathbf{Q}(\theta_1)\| \ge 2\|\mathbf{Q}(\theta_0)\| \ge 4n_0$, we continue with policy $\mathcal{U}$ until the corresponding time θ_2 and check whether $\|\mathbf{Q}(\theta_2)\| \ge 2\|\mathbf{Q}(\theta_1)\| \ge 8n_0$, and so on. Either this process continues indefinitely or for some time instance θ_i, we get $\|\mathbf{Q}(\theta_i)\| < 2\|\mathbf{Q}(\theta_{i-1})\|$. Set $\theta_{-1} = 0$ by convention. Let $\mathcal{E}_m, m = 0, 1, \dots,$ denote the event $\|\mathbf{Q}(\theta_i)\| \ge 2\|\mathbf{Q}(\theta_{i-1})\|$ and

$$\inf_{\theta_{i-1} \le t \le \theta_i} \|\mathbf{Q}(t)\| \ge \frac{\|\mathbf{Q}(\theta_{i-1})\|}{4} \max\left(\frac{\gamma}{C}, 1\right), \tag{91}$$

for all $i \le m$. In particular, the event implies $\|\mathbf{Q}(\theta_m)\| \ge 2^{m+1}n_0 \ge n_0$. Let $\mathcal{E}^1 = \bigcap_m \mathcal{E}_m$, that is, $\mathcal{E}^1$ implies that the process of exceeding the bounds continues indefinitely. We now show that $\mathbb{P}\{\mathcal{E}^1\} \ge \alpha > 0$, where α depends only on the parameters of the model and on n_0 and γ (and is independent for example from the components z_1, z_2). By (86) and (87) of Theorem 14, the probability of the event $\mathcal{E}^1 \equiv \bigcap_{m=0}^{\infty} \mathcal{E}_i$ is at least

$$\begin{aligned} 1 - 2\sum_{m=0}^{\infty} O\bigl(\exp(-\Theta(2^m\|q\|))\bigr) &> 1 - \sum_{m=0}^{\infty} O\bigl(\exp(-\Theta((m+1)\|q\|))\bigr) \\ &= 1 - \frac{O(\exp(-\Theta(\|q\|)))}{1 - O(\exp(-\Theta(\|q\|)))} \\ &> 1 - O\bigl(e^{-\Theta(n_0)}\bigr). \end{aligned}$$

We take n_0 sufficiently large so that $\alpha \equiv 1 - O(e^{-\Theta(n_0)}) > 0$. The parameters hidden in $\Theta(\cdot)$ depend only on the parameters of the model (including the large deviations parameters V, L) and γ. Thus, the probability of $\mathcal{E}^1$ is positive (and, in fact, is close to unity), provided that n_0 is sufficiently large.

Next, we show that the event $\mathcal{E}^1$ implies

$$\liminf_t \frac{\|\mathbf{Q}(t)\|}{t} \geq \frac{\max(\gamma/C, 1)}{8\max(1, 3/\gamma)} > 0. \tag{92}$$

We first show by induction in m that event $\mathcal{E}^1$ implies $\|\mathbf{Q}(\theta_m)\|/\theta_m \geq 1/\theta$ for all $m \geq 1$. When $m = 1$, the ratio is at least $2\|q\|/(\theta\|q\|) > 1/\theta$. Suppose the assertion holds for $i = 1, 2, \dots, m-1$. Note that $\theta_m = \theta_{m-1} + \theta\|\mathbf{Q}(\theta_{m-1})\|$ and by $\mathcal{E}^1$, $\|\mathbf{Q}(\theta_m)\| \geq 2\|\mathbf{Q}(\theta_{m-1})\|$. Therefore,

$$\begin{aligned}\frac{\|\mathbf{Q}(\theta_m)\|}{\theta_m} &\geq \frac{2\|\mathbf{Q}(\theta_{m-1})\|}{\theta_{m-1} + \theta\|\mathbf{Q}(\theta_{m-1})\|} \\ &= \frac{2\|\mathbf{Q}(\theta_{m-1})\|/\theta_{m-1}}{1 + \theta\|\mathbf{Q}(\theta_{m-1})\|/\theta_{m-1}}.\end{aligned}$$

But by the inductive assumption, $\|\mathbf{Q}(\theta_{m-1})\|/\theta_{m-1} \geq 1/\theta$. This immediately implies that the expression above is also at least $1/\theta$, and the induction is completed. Now for every $t \geq \theta_0 = \theta\|q\|$, we find θ_m such that $\theta_{m-1} \leq t < \theta_m$. Using (91), we obtain

$$\begin{aligned}\frac{\|\mathbf{Q}(t)\|}{t} &\geq \frac{\|\mathbf{Q}(\theta_{m-1})\|}{4\theta_m} \max\left(\frac{\gamma}{C}, 1\right) \\ &= \frac{1}{4(\theta_{m-1}/\|\mathbf{Q}(\theta_{m-1})\|) + 4\theta} \max\left(\frac{\gamma}{C}, 1\right) \\ &\geq \frac{\max(\gamma/C, 1)}{8\theta} > 0,\end{aligned}$$

where the last inequality follows since $\|\mathbf{Q}(\theta_{m-1})\|/\theta_{m-1} \geq 1/\theta$. This shows (92).

Now, suppose the event $\mathcal{E}_i$ fails to occur at some θ_i, and thus $\mathcal{E}^1$ does not occur. We then "restart" the process of attempting to obtain an infinite sequence of points θ_i with the properties outlined above. Let us call $\mathcal{E}^2$ the event that the sequence is obtained after restarting the process again as follows. At that time at which $\mathcal{E}_i$ fails, we switch to any nonidling nonpreemptive scheduling policy. Applying Assumption C, with probability one, there exists a time τ_1 for which $\|\mathbf{Q}(\tau_1)\| \geq n_0$. Note that it is possible that $\tau_1 = \theta_i$. We apply the policy $\mathcal{U}$ starting from time τ_1. Repeating the argument for $\mathcal{E}^1$, with probability greater than α, we obtain a new infinite sequence of time instances θ_i' such that $\|\mathbf{Q}(\theta_{i+1}')\| \geq 2\|\mathbf{Q}(\theta_i')\|$, that is, $\mathcal{E}^2$ occurs. If $\mathcal{E}^2$ does not occur, we again restart the process. Finally, the probability of eventually obtaining a sequence of points θ_i with the stated properties is given by $\mathbb{P}(\mathcal{E}) \equiv \mathbb{P}(\bigcup_{k=1}^{\infty} \mathcal{E}^k)$, where the $\mathcal{E}^k$ are defined as above in the natural way. Since the probability of each event $\mathcal{E}^k$ is bounded below by α, and this lower bound on probability does not depend on whether or not the other events occur, the probability of $\mathcal{E}$ is one.

Finally, we show below that the event $\mathcal{E}$ implies (22), that is,

$$\liminf_t \frac{\|\mathbf{Q}(t)\|}{t} \geq \frac{\max(\gamma/C, 1)}{8\max(1, 3/\gamma)} > 0. \tag{93}$$

Let k be the smallest integer for which $\mathcal{E}_k$ occurs. Denote by T_k the time corresponding to the beginning of this event. Fix any state (q, z_1, z_2) and $t_0 > 0$ and condition on $T_k = t_0, \mathbf{Q}(t_0) = q$. Applying (92) to the event $\mathcal{E}_k$, we obtain

$$\liminf_t \frac{\|\mathbf{Q}(t)\|}{t_0 + t} \geq \frac{\max(\gamma/C, 1)}{8\max(1, 3/\gamma)} > 0. \tag{94}$$

Since t_0 is fixed, the lower bound (93) holds as well. Integrating over the choices of (q, z_1, z_2) and t_0, we complete the proof of the theorem. □

5. Conclusions and further work.

The present work leaves many interesting questions open. The most immediate one is whether the result connecting global weak stability and rate stability holds for networks with any number of stations. One way to prove this conjecture would be to establish the Finite Decomposition Property for fluid networks with more than two stations. Of course, the question of whether (strong) global stability of the fluid model is equivalent to positive Harris recurrence remains open even for networks with two stations. There the difficulty lies in being able to analyze the dynamics of the stochastic network at the critical regime $\rho^* = 1$. Finally, we mention that our assumption that interarrival and service times are i.i.d. is used to simplify the exposition and our result should hold for networks with more general primitives as long as the associated processes satisfy appropriate large deviations bounds.

APPENDIX

PROOF OF LEMMA 1. We begin by proving (9). The proof of (10) is then derived using (9). Our method uses the standard derivation of LD upper bounds on i.i.d. sequences.

Part I. Let us fix arbitrary $\varepsilon > 0$ and $\theta > 0$. Then we note the following hold for all $n \geq 1$:

$$\begin{aligned}\mathbb{P}\Biggl\{\sum_{1\leq i\leq n} Z_i \geq n\alpha + n\varepsilon + z \,\Bigg|\, Z_1 \geq z\Biggr\} &= \mathbb{P}\{e^{\theta \sum_{1\leq i\leq n} Z_i} \geq e^{\theta(n\alpha+n\varepsilon+z)} \mid Z_1 \geq z\} \\ &\leq \frac{\mathbb{E}[e^{\theta(Z_1-z)} \mid Z_1 \geq z](\mathbb{E}[e^{\theta Z_2}])^{n-1}}{e^{n\theta(\alpha+\varepsilon)}} \\ &\leq \frac{F(\theta)(\mathbb{E}[e^{\theta Z_2}])^n}{e^{n\theta(\alpha+\varepsilon)}},\end{aligned}$$

where we use $z \geq 0$ and $\mathbb{E}[e^{\theta Z_2}] \geq 1$. It is a standard result in large deviations theory [14] that $\mathbb{E}[e^{\theta Z_2}]/e^{\theta(\alpha+\varepsilon)} \equiv e^{-L(\varepsilon)} < 1$ for some value of $\theta = \theta(\varepsilon) \in [0, \theta_0]$

as long as $\mathbb{E}[e^{\theta Z_2}]$ is finite on $[0,\theta_0]$. Thus, our tail probability is at most $F(\theta(\varepsilon))e^{-L(\varepsilon)n}$. We fix a suitable θ and take $V = F(\theta)$.

We now prove a complimentary bound. Again fix arbitrary $\varepsilon > 0$ and $\theta > 0$,

$$\begin{aligned}\mathbb{P}\left\{\sum_{1\le i\le n} Z_i \le n\alpha - n\varepsilon + z \,\middle|\, Z_1 \ge z\right\} &= \mathbb{P}\{e^{-\theta\sum_{1\le i\le n} Z_i} \ge e^{-\theta n\alpha+\theta n\varepsilon-\theta z} \mid Z_1 \ge z\}\\ &\le \frac{\mathbb{E}[e^{-\theta(Z_1-z)} \mid Z_1 \ge z](\mathbb{E}[e^{-\theta Z_2}])^{n-1}}{e^{-n\theta\alpha+n\theta\varepsilon}}\\ &\le \frac{(\mathbb{E}[e^{-\theta Z_2}])^{n-1}}{e^{-(n-1)\theta(\alpha-\varepsilon)}}e^{\theta\alpha-\theta\varepsilon},\end{aligned}$$

where we use $\mathbb{E}[e^{-\theta(Z_1-z)} \mid Z_1 \ge z] \le 1$. Again we use a standard result in large deviations theory [14] stating that $\mathbb{E}[e^{-\theta Z_2}]/e^{-(\theta\alpha-\theta\varepsilon)} \equiv e^{-L(\varepsilon)} < 1$ for some value of $\theta = \theta(\varepsilon) \in [0,\theta_0]$. We take $V = e^{L(\varepsilon)+\theta\alpha-\theta\varepsilon}$. This proves (9).

Part II. We now prove (10). Consider a fixed, but arbitrary $\varepsilon > 0$.

We first obtain a bound which is valid for all $t \ge 1$. Note that (9) easily implies two one-sided versions of the LD inequality. Applying one such one-sided version of (9) with $n = \lceil t/\alpha + \varepsilon t\rceil \ge 1$ and $\tilde{\varepsilon} = \alpha^2\varepsilon/(1+\alpha\varepsilon+\alpha) > 0$, then there exist $\tilde{L}, V_1 > 0$ such that

$$\mathbb{P}\left\{\sum_{i=1}^{\lceil t/\alpha+\varepsilon t\rceil} Z_i - z \le \alpha\lceil t/\alpha+\varepsilon t\rceil - \tilde{\varepsilon}\lceil t/\alpha+\varepsilon t\rceil \,\middle|\, Z_1 \ge z\right\} \le V_1 e^{-\tilde{L}\lceil t/\alpha+\varepsilon t\rceil},$$

for all $t \ge 0$. Next since $\lceil t/\alpha + \varepsilon t\rceil \ge t/\alpha + \varepsilon t$ and for $t \ge 1$, $\lceil t/\alpha + \varepsilon t\rceil \le t/\alpha + \varepsilon t + t$, we have

$$\mathbb{P}\left\{\sum_{i=1}^{\lceil t/\alpha+\varepsilon t\rceil} Z_i - z \le \alpha[t/\alpha+\varepsilon t] - \tilde{\varepsilon}[t/\alpha+\varepsilon t+t] \,\middle|\, Z_1 \ge z\right\} \le V_1 e^{-\tilde{L}\lceil t/\alpha+\varepsilon t\rceil},$$

for all $t \ge 1$. Multiplying through inside the probability yields

$$\mathbb{P}\left\{\sum_{i=1}^{\lceil t/\alpha+\varepsilon t\rceil} Z_i - z \le t + \alpha\varepsilon t - \alpha\varepsilon t \,\middle|\, Z_1 \ge z\right\} \le V_1 e^{-\tilde{L}\lceil t/\alpha+\varepsilon t\rceil}.$$

Further simplification gives

$$\mathbb{P}\left\{\sum_{i=1}^{\lceil t/\alpha+\varepsilon t\rceil} Z_i \le t + z \,\middle|\, Z_1 \ge z\right\} \le V_1 e^{-\tilde{L}\lceil t/\alpha+\varepsilon t\rceil},$$

for all $t \ge 1$. Using the duality relationship between a counting process and its increments, the above implies

$$\mathbb{P}\{N(t+z) \ge t/\alpha + \varepsilon t \mid Z_1 \ge z\} \le V_1 e^{-\tilde{L}\lceil t/\alpha+\varepsilon t\rceil}. \tag{95}$$

Finally, setting $L_1 = \tilde{L}(1/\alpha + \varepsilon)$, we have

$$V_1 e^{-\tilde{L}\lceil t/\alpha+\varepsilon t\rceil} \le V_1 e^{-L_1 t},$$

for all $t \ge 1$. Hence, we can rewrite (95) as

$$\mathbb{P}\{N(t+z) \ge t/\alpha + \varepsilon t \mid Z_1 \ge z\} \le V_1 e^{-L_1 t}, \tag{96}$$

for all $t \ge 1$.

For $t < 1$ and any $V_2 > 1$, note that $V_2 \cdot \exp(-L_2 t) \ge 1$ if we set $L_2 \equiv \ln V_2 > 0$. Hence,

$$\mathbb{P}\{N(t+z) \ge t/\alpha + \varepsilon t \mid Z_1 \ge z\} \le V_2 e^{-L_2 t} \tag{97}$$

holds trivially for all $t < 1$ with such a V_2 and L_2. Finally, setting $V_3 \equiv \max\{V_1, V_2\}$ and $L_3 \equiv \min\{L_1, L_2\}$, then combining with (96) and (97), we obtain

$$\mathbb{P}\{N(t+z) \ge t/\alpha + \varepsilon t \mid Z_1 \ge z\} \le V_3 e^{-L_3 t},$$

for all $t \ge 0$.

This proves one side of the inequality in (10). The other direction is proved by an exactly analogous argument. The final result is then obtained by combining the two directions, applying Boole's inequality and again using appropriate constants B and L. □

Acknowledgments. We would like to thank several anonymous referees for helpful technical and expositional comments.

REFERENCES

[1] BERTSIMAS, D., GAMARNIK, D. and TSITSIKLIS, J. N. (1996). Stability conditions for multiclass fluid queueing networks. *IEEE Trans. Automat. Control* **41** 1618–1631.

[2] BRAMSON, M. (1999). A stable queueing network with unstable fluid model. *Ann. Appl. Probab.* **9** 818–853.

[3] BRAMSON, M. (2001). Stability of earliest-due-date, first-served queueing networks. *Queueing Syst. Theory Appl.* **39** 79–102.

[4] CHEN, H. (1995). Fluid approximations and stability of multiclass queueing networks I: Work-conserving disciplines. *Ann. Appl. Probab.* **5** 637–665.

[5] CHEN, H. and ZHANG, H. (1997). Stability of multiclass queueing networks under FIFO service discipline. *Math. Oper. Res.* **22** 691–725.

[6] COFFMAN, E. and STOLYAR, A. (2001). Bandwidth packing. Average-case analysis of algorithms. *Algorithmica* **29** 70–88.

[7] DAI, J. G. (1995). On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* **5** 49–77.

[8] DAI, J. G. (1996). A fluid-limit model criterion for instability of multiclass queueing networks. *Ann. Appl. Probab.* **6** 751–757.

[9] DAI, J. G. (1999). Stability of fluid and stochastic processing networks. In *MaPhySto Miscellanea Publication* **9**. Centre for Mathematical Physics and Stochastics.

[10] DAI, J. G., HASENBEIN, J. J. and VANDEVATE, J. H. (2004). Stability and instability of a two-station queueing network. *Ann. Appl. Probab.* **14** 326–377.

[11] DAI, J. G. and JENNINGS, O. B. (2004). Stabilizing queueing networks with setups. *Math. Oper. Res.* **29** 891–922.

[12] DAI, J. G. and VANDEVATE, J. (2000). The stability of two-station multitype fluid networks. *Oper. Res.* **48** 721–744.

[13] DANTZER, J.-F. and ROBERT, P. (2002). Fluid limits of string valued Markov processes. *Ann. Appl. Probab.* **12** 860–889.

[14] DEMBO, A. and ZEITOUNI, O. (1998). *Large Deviations Techniques and Applications*. Springer, New York.

[15] GAMARNIK, D. (2002). On deciding stability of constrained homogeneous random walks and queueing systems. *Math. Oper. Res.* **27** 272–293.

[16] KUMAR, S. and KUMAR, P. R. (1996). Fluctuation smoothing policies are stable for stochastic reentrant lines. *Discrete Event Dyn. Syst.* **6** 361–370.

[17] MEYN, S. P. (1995). Transience of multiclass queueing networks via fluid limit models. *Ann. Appl. Probab.* **5** 946–957.

[18] MEYN, S. P. (2001). Sequencing and routing in multiclass queueing networks. Part I: Feedback regulation. *SIAM J. Control Optim.* **40** 741–776.

[19] PUHALSKII, A. A. and RYBKO, A. N. (2000). Nonergodicity of a queueing network under nonstability of its fluid model. *Probl. Inf. Transm.* **36** 23–41.

[20] RYBKO, A. N. and STOLYAR, A. L. (1992). Ergodicity of stochastic processes describing the operation of open queueing networks. *Probl. Inf. Transm.* **28** 199–220.

[21] STOLYAR, A. (1995). On the stability of multiclass queueing networks: A relaxed sufficient condition via limiting fluid processes. *Markov Process. Related Fields* **1** 491–512.

IBM T. J. WATSON RESEARCH CENTER
YORKTOWN HEIGHTS, NEW YORK 10598
USA
E-MAIL: gamarnik@watson.ibm.com

DEPARTMENT OF MECHANICAL ENGINEERING
UNIVERSITY OF TEXAS AT AUSTIN
1 UNIVERSITY STATION AUSTIN, TEXAS 78712-0292
USA
E-MAIL: jhas@mail.utexas.edu