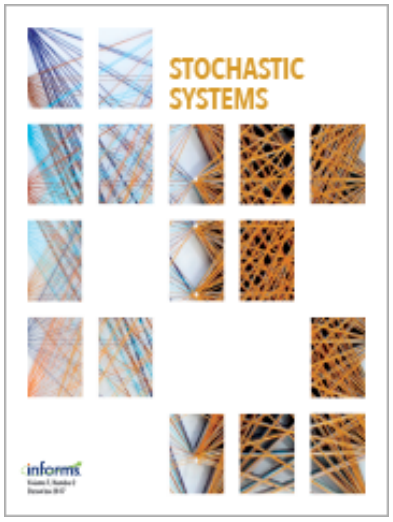


Stochastic Systems

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

Jumping Fluid Models and Delay Stability of Max-Weight Dynamics Under Heavy-Tailed Traffic

Arsalan Sharifnassab, John Tsitsiklis

To cite this article:
Arsalan Sharifnassab, John Tsitsiklis (2023) Jumping Fluid Models and Delay Stability of Max-Weight Dynamics Under Heavy-Tailed Traffic. Stochastic Systems 13(4):399-437. https://doi.org/10.1287/stsy.2023.0110

https://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 13, No. 4, December 2023, pp. 399–437
ISSN 1946-5238 (online)

Jumping Fluid Models and Delay Stability of Max-Weight Dynamics Under Heavy-Tailed Traffic

Arsalan Sharifnassab,[a,*] John Tsitsiklis[b]

[a] Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2R3, Canada; [b] Laboratory for Information and Decision Systems, Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139
*Corresponding author
Contact: sharifna@ualberta.ca, https://orcid.org/0000-0002-3910-2878 (AS); jnt@mit.edu, https://orcid.org/0000-0003-2658-8239 (JT)

Received: November 16, 2021
Revised: February 24, 2023
Accepted: March 6, 2023
Published Online in Articles in Advance: April 14, 2023

https://doi.org/10.1287/stsy.2023.0110

Abstract. We say that a random variable is *light-tailed* if moments of order $2+\epsilon$ are finite for some $\epsilon > 0$; otherwise, we say that it is *heavy-tailed*. We study queueing networks that operate under the max-weight scheduling policy for the case in which some queues receive heavy-tailed and some receive light-tailed traffic. Queues with light-tailed arrivals are often delay stable (that is, expected queue sizes are uniformly bounded over time) but can also become delay unstable because of resource sharing with other queues that receive heavy-tailed arrivals. Within this context and for any given "tail exponents" of the input traffic, we develop a necessary and sufficient condition under which a queue is robustly delay stable, in terms of *jumping fluid* models—an extension of traditional fluid models that allows for jumps along coordinates associated with heavy-tailed flows. Our result elucidates the precise mechanism that leads to delay instability through a coordination of multiple abnormally large arrivals at possibly different times and queues and settles an earlier open question on the sufficiency of a particular fluid-based criterion. Finally, we explore the power of Lyapunov functions in the study of delay stability.

Keywords: queueing networks • network scheduling • max-weight algorithm • heavy-tailed traffic • fluid model

1. Introduction

We say that a random variable is *light-tailed* if moments of order $2+\epsilon$ are finite for some $\epsilon > 0$; otherwise, we say that it is *heavy-tailed*. We study queueing networks that operate under the max-weight (MW) scheduling policy for the case in which some queues receive heavy-tailed traffic, whereas some other queues receive light-tailed traffic; our motivation stems from the fact that heavy-tailed processes are often natural models of the inputs to computer and communications networks (Foss et al. 2011). Queues that receive heavy-tailed traffic are naturally delay unstable, that is, they incur infinite expected delay as an immediate consequence of the Pollaczek–Khintchine formula. However, it is also known that, because of the relatively complex max-weight dynamics, some of the queues that receive light-tailed traffic may also end up delay unstable.[1] Our aim is to develop conditions that determine whether any particular queue is delay stable or not.

This problem is studied extensively (Markakis 2013; Markakis et al. 2014, 2018; Nair et al. 2015), and a necessary condition for delay stability is given in Markakis et al. (2014, 2018). In particular, Markakis et al. (2018) consider the associated fluid model, initialized at zero, except for a positive initial condition at some queue that receives heavy-tailed arrivals. If another queue happens to eventually become positive under that fluid model, one can then conclude that the latter queue is delay unstable. This result leads to the natural question whether this condition is also sufficient, that is, whether delay stability is guaranteed when any such fluid trajectory (with a positive initialization at any single one of the queues that receive heavy-tailed traffic) keeps the queue of interest at zero level. Such a sufficiency result might appear plausible because, in models involving heavy-tailed random variables, large fluctuations are often the consequence of a single abnormally large value in the underlying heavy-tailed random variables (Pakes 1975, Veraverbeke 1977, Anantharam 1989, Asmussen 1996, Durrett 1980, Foss et al. 2011).

1.1. Our Contributions

We start in Section 3 by showing that the aforementioned possible sufficiency result does not hold. We accomplish this by providing a fairly simple example in which a large arrival at any single heavy-tailed queue does not cause a certain queue of interest to grow, but a combination of two large arrivals, at two different heavy-tailed queues, can result in large backlogs at the queue of interest. We also provide necessary and sufficient conditions for delay stability in that particular example, and then provide intuition for a possible more general result. Interestingly, delay instability manifests itself only when the tail exponents (defined precisely in Section 2.2) of the heavy-tailed arrival processes lie in a specific range. As a consequence, we need to take these exponents explicitly into account, something that traditional fluid models cannot do.

We then generalize, by developing general and tight (necessary and sufficient) conditions for delay stability, in terms of deterministic fluid-like models in which there can be multiple jumps (in different queues and possibly at different times) at the heavy-tailed queues.

Our conditions are not easy to test computationally, but this seems unavoidable: because the conditions are necessary and sufficient, the complexity of testing them reflects the intrinsic complexity of testing delay stability. On the positive side, our conditions

a. Provide a conceptual understanding of the mechanism that results in delay instability.
b. Can be checked in special cases, for instance, for the example in Section 3; see Section 5.3.

On the technical side, our general conditions involve a small but technically crucial reformulation of the delay stability problem. To understand the underlying issue, note that fluid models do not always lead to definite conclusions when the underlying system is marginally stable but are generally effective when used to analyze "robust" properties. For this reason, we consider a network with given nominal arrival rates and focus on robust stability. Namely, we ask whether a certain queue is delay stable for all arrival processes with given tail exponents and for all (possibly time-varying) arrival rates that lie in some open ball around the nominal ones. When the problem is framed that way, definitive necessary and sufficient conditions for (robust) delay stability become possible. Furthermore, with this formulation, it is only the nominal arrival rates and the tail exponents that matter as opposed to the details of the distribution of the input traffic.

Finally, earlier works Markakis (2013) and Markakis et al. (2018) show that Lyapunov functions with certain structural properties can be used to certify delay stability. But it was not known whether this methodology is "complete," that is, whether delay stability can always be established through a suitable Lyapunov function. Our results make progress in this direction, for the case of "very heavy" tails, that is, when there exists some $\gamma > 0$ for which arriving traffic moments of order $1 + \gamma$ are infinite but γ can be arbitrarily small. For this regime, we derive some necessary and sufficient conditions for delay stability: we show that a Lyapunov function of a special kind can be used to certify delay stability together with a partial converse.

1.2. Related Works on Multiple Big Jumps

As already mentioned, in several systems that involve heavy-tailed random variables, large fluctuations are the consequence of a single large jump in the underlying heavy-tailed variables. However, this is not always the case. There are known systems that have small fluctuations under a single large jump, and large deviations can only arise as a consequence of multiple jumps in heavy-tailed random variables with suitable timing. Early examples are studied in Jelenković and Momčilović (2003) and Zwart et al. (2004) for a single-buffer system with multiple on/off input processes with heavy-tailed periods. Furthermore, Foss and Korshunov (2012) provide conditions on the finiteness of moments of queue sizes in a multiserver G/G/s queue and show the special effects caused by multiple jumps. Perhaps the closest work to ours that demonstrates the necessity of multiple big jumps is Chen et al. (2019), who propose a rare-event simulation technique to estimate the probability of rare events in stochastic systems that receive heavy-tailed inputs. As an application, they consider a queuing network with fixed fluid services and fixed routing in which each queue receives independent heavy- or light-tailed exogenous arrivals. They characterized the tail exponent of the queues in terms of a knapsack-type constraint that involves the sum of numbers of large jumps in the different inputs that are required to make the queue of interest large, in which the sum is weighted by the tail exponents of the corresponding inputs. Such a knapsack-type constraint also appears in our bounds, and the tail exponent fully determines the delay stability/instability of the queue of interest. Thus, our results parallel those in Chen et al. (2019) but for a different setting. The main differences are that our network operates under the more complex max-weight scheduler, and our assumptions allow for non–independent and identically distributed (i.i.d.) arrival processes. Furthermore, our proof techniques are very different from those in Chen et al. (2019); it would be interesting to explore whether the techniques in Chen et al. (2019) can be used to obtain alternative proofs for our setting (Theorem 1).

1.3. Outline

The rest of the paper is organized as follows. We start in the next section with the details of our model and some definitions. In Section 3, we discuss an example that provides insights on the ways that arrival rates and tail exponents affect delay stability and demonstrate that criteria based on traditional fluid models are inadequate for the purpose of deciding delay stability. In Section 4, we introduce a fluid-like model, which we call the *jumping fluid (JF) model* and underlies our main result, the necessary and sufficient conditions for (robust) delay stability that we present in Section 5. In Section 6, we study the power of Lyapunov functions for our problem. In Sections 7–9, we provide the proofs of our results. As the proofs are quite involved, they are presented as a sequence of lemmas with the proofs of the lemmas provided in Appendices B and C. We discuss the results and directions for future research in Section 10. Finally, in Appendix A, we explore alternative definitions of robust stability and corresponding variants of our jumping fluid conditions and explain why they are unlikely to yield sharp necessary and sufficient conditions.

1.4. Notation

We collect here some notational conventions to be used throughout the paper. We use boldface symbols to denote vectors and ordinary font to denote scalars. For any vector $\mathbf{v}$, we use v_i to denote its ith component and $|\mathbf{v}|$ to denote the sum $|v_1|+\cdots+|v_n|$. We also use the notation $[\mathbf{v}]^+$ to denote the vector with components $\max\{0, v_i\}$. Finally, we let $\mathbf{e}_j$ stand for the jth unit vector.

We use $\mathbb{R}_+$ and $\mathbb{Z}_+$ to denote the sets of nonnegative reals and nonnegative integers, respectively. Furthermore, for a vector $\mathbf{v}$, we write $\mathbf{v} \geq \mathbf{0}$ (respectively, $\mathbf{v} > \mathbf{0}$) to indicate that all components are nonnegative (positive). For any set S, we denote its convex hull by $\text{conv}(S)$.

Throughout, $\|\cdot\|$ stands for the Euclidean norm. Sometimes, we use the alternative notation $d(\mathbf{x}, \mathbf{y})$ in place of $\|\mathbf{x}-\mathbf{y}\|$. We also let $d(\mathbf{x}, S)$ be the distance of a vector $\mathbf{x}$ from a set S, that is, $d(\mathbf{x}, S) = \inf_{\mathbf{y}\in S}\|\mathbf{x}-\mathbf{y}\|$. We finally use $\mathbb{1}(\cdot)$ to denote the indicator function and log to denote the natural logarithm.

For any time function $\mathbf{x}(\cdot)$ that is right-continuous with left limits, $(d\mathbf{x}/dt)(t)$ or $\dot{\mathbf{x}}(t)$ stand for the right derivative of $\mathbf{x}(t)$ with the implicit assumption that it exists, and $\mathbf{x}(t^-)$ stands for $\lim_{\tau\uparrow t}\mathbf{x}(\tau)$.

2. The Model

2.1. Network Model and the Max-Weight Policy

We consider a switched network that operates in discrete time. For simplicity and ease of presentation, we restrict ourselves to single-hop networks. However, our results are easily generalized to multihop networks of the type considered in Sharifnassab et al. (2020).

The network consists of ℓ queues that buffer incoming packets (or jobs). For any $t \in \mathbb{Z}_+$, we let $\mathbf{Q}(t)$ be a nonnegative vector whose jth component is the length of the jth queue at time t. Packets arrive to the queues according to a nonnegative stochastic vector arrival process, $\mathbf{A}(\cdot)$. In particular, $A_j(t)$ stands for the exogenous arrival to the jth queue at time t. We assume that the random variables $A_j(t)$, for different j and t, are independent. We refer to $\mathbb{E}[\mathbf{A}(t)]$ as the arrival rate vector at time t.

At each time t, the amount of service received by the queues is a nonnegative vector $\boldsymbol{\mu}(t)$, which is chosen by a scheduler from a finite set $\mathcal{M}$ of possible service vectors. The queue lengths then evolve according to

$$\mathbf{Q}(t+1) = [\mathbf{Q}(t) - \boldsymbol{\mu}(t)]^+ + \mathbf{A}(t). \tag{1}$$

As in Sharifnassab et al. (2020), we assume throughout the paper that, for any $\boldsymbol{\mu} \in \mathcal{M}$, the set $\mathcal{M}$ also contains all vectors that result from setting some entries of $\boldsymbol{\mu}$ to zero. This assumption is naturally valid in most contexts.

We focus exclusively on the popular MW scheduling policy,

$$\boldsymbol{\mu}(t) \in \operatorname*{argmax}_{\boldsymbol{\nu}\in\mathcal{M}} \boldsymbol{\nu}^T\mathbf{Q}(t), \tag{2}$$

which is known to have favorable stability properties (Tassiulas and Ephremides 1992): whenever there exists a policy under which the queues remain stable, MW results in stable queues. More specifically, let us consider the set $\overline{\mathcal{M}}$, defined as the convex hull, $\text{conv}(\mathcal{M})$, of the set of all possible service vectors, which is the so-called *capacity region* of the network. For the case of i.i.d. arrivals and under common stochastic assumptions, it is known that, if the arrival rate vector lies in the interior of the capacity region, then MW results in stable queues; conversely, if the arrival rate vector lies outside the capacity region, the queues are unstable under every scheduling policy (Tassiulas and Ephremides 1992).

2.2. Light- and Heavy-Tailed Arrivals

In this section, we present some definitions related to the tails of the arrival process distributions.

To any nonnegative random variable X, we associate a *tail exponent*, defined as the value of γ at which $\mathbb{E}[X^{1+\gamma}]$ switches from finite to infinite:

$$\gamma^* = \sup\{\gamma : \mathbb{E}[X^{1+\gamma}] < \infty\}. \tag{3}$$

As an example, consider a continuous random variable whose probability density function $f(\cdot)$ satisfies

$$c \cdot x^{-(2+\gamma)} \le f(x) \le \log^k x \cdot x^{-(2+\gamma)}, \qquad \forall x \ge x_0, \tag{4}$$

where c, γ, k, and x_0 are positive constants. Such a distribution has a tail exponent equal to γ.

For an i.i.d. arrival process, a tail exponent is unambiguously defined as the tail exponent of the marginal distribution at an arbitrary time. However, once we bring robustness into the picture, we are led to consider arrival processes with nonidentically distributed $\mathbf{A}(t)$. To any arrival process $A_j(\cdot)$, we associate a tail exponent, γ_j, defined as the largest value of γ such that $A_j(t)$ is dominated by some nonnegative random variable X with tail exponent γ for all times t:

$$\gamma_j = \sup\{\gamma : \text{there exists a r.v. } X \ge 0 \quad \text{s.t.} \quad \begin{array}{l} X \text{ dominates } A_j(t) \text{ for all } t \ge 0, \\ \text{and } \mathbb{E}[X^{1+\gamma}] < \infty\}. \end{array} \tag{5}$$

Here, the term "dominates" refers to stochastic dominance: a random variables X dominates a random variable Y if $\mathbb{P}(X > a) \ge \mathbb{P}(Y > a)$ for all $a \in \mathbb{R}$. We say that $A_j(\cdot)$ is heavy-tailed if $\gamma_j \le 1$ and light-tailed otherwise. Note that this definition of a heavy-tailed process is aimed to capture the boundedness of variance of the input distribution and differs from the conventional definition of heavy-tailed random variables, which requires subexponential decay of tail probabilities (Nair et al. 2020). The tail behavior of the different arrival processes is summarized by the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_\ell)$.

We note that, as long as $\mathbb{E}[A_j(t)] \le \overline{\mu}$ for some constant $\overline{\mu}$ and for all times t, the tail exponent γ_j is well-defined and lies in the range $[0, \infty]$. Indeed, suppose that $\lambda_j(t) = \mathbb{E}[A_j(t)] \le \overline{\mu}$ for some finite constant $\overline{\mu}$ and for all t. Consider a random variable X with probability density function $f_X(x) = (\overline{\mu}/x^2)\,\mathbb{1}(x \ge \overline{\mu})$. The Markov inequality implies that

$$\mathbb{P}(A_j(t) > a) \le \min\{1, \mathbb{E}[A_j(t)]/a\} \le \min\{1, \overline{\mu}/a\} = \mathbb{P}(X > a), \qquad \forall a > 0.$$

In particular, X dominates $A_j(t)$ for all t. Furthermore, $\mathbb{E}[X^{1+\gamma}] < \infty$, for every $\gamma < 0$. Thus, γ_j is at least as large as any negative number, which implies that $\gamma_j \ge 0$.

Conversely, if $\gamma_j > 0$, then $\mathbb{E}[A_j(t)]$ is finite and bounded as a function of t. We finally note that γ_j may be infinite; this is the case for bounded or, more generally, exponential-type distributions.

2.3. Robust Delay Stability

As already mentioned, we are interested in the question of delay stability under the MW policy in the presence of heavy-tailed arrivals. Given a set of arrival processes, we say that queue m is *delay stable* if, starting from $\mathbf{Q}(0) = 0$, we have $\sup_t \mathbb{E}[Q_m(t)] < \infty$.

The arrival rates and the tail exponents do not provide enough information to decide whether we have delay stability or even stability. For example, there is sometimes indeterminacy on the boundary of the capacity region. Furthermore, a tail exponent of $\gamma_j = 1$ is compatible with $\mathbb{E}[A_j^2(t)]$ being either finite or infinite, and this may be critical as far as delay stability is concerned (cf. the Pollaczek–Khintchine formula). These difficulties, all related to indeterminacy at certain boundaries, can be circumvented by focusing on a robust version of delay stability that incorporates two distinct elements.

a. We require delay stability for all arrival process distributions with given tail exponents. This allows us to focus on conditions that involve the tail exponents and ignore other details of these distributions.

b. We require delay stability for all arrival rates, possibly time-varying, in the vicinity of a given nominal rate. This allows us to avoid issues of indeterminacy when the arrival rate lies at a threshold between delay stability and instability.

Definition 1 (Robust Delay Stability). Let us fix a network with ℓ nodes and a set $\mathcal{M}$ of possible service vectors. We consider a tail exponent vector $\boldsymbol{\gamma}$ with components in $[0, \infty]$ and a nominal arrival rate vector $\boldsymbol{\lambda}^* \ge \mathbf{0}$.

a. Given some $\delta \ge 0$, an arrival process $\mathbf{A}(\cdot)$ belongs to the class $\mathcal{A}_\delta(\boldsymbol{\gamma}; \boldsymbol{\lambda}^*)$ if
   i. The random variables $A_j(t)$, for different j and t, are independent and have finite means.
   ii. For every j, the tail exponent of the process $A_j(\cdot)$ is γ_j.
   iii. For every $t \ge 0$, we have $\|\mathbb{E}[\mathbf{A}(t)] - \boldsymbol{\lambda}^*\| \le \delta$.

b. For $m \in \{1, \ldots, \ell\}$, we say that queue m is robustly delay stable (RDS) if there exists some $\delta > 0$ such that queue m is delay stable for all arrival processes in the class $\mathcal{A}_\delta(\boldsymbol{\gamma}; \boldsymbol{\lambda}^*)$.

3. A Counterexample and the Insufficiency of Fluid Models

In this section, we discuss a simple example with the following features:

a. Similar to existing examples, heavy-tailed arrivals to some queues can cause delay instability at a queue that receives light-tailed traffic.

b. Tight criteria for delay instability need to take into account the values of the tail exponents. In particular, criteria that are based on traditional fluid models cannot be conclusive because they do not involve the tail exponents.

c. Delay instability may emerge from coordinated large arrivals at multiple heavy-tailed queues.

Consider the three-queue network in Figure 1. The set of possible service vectors $\mathcal{M}$ is such that, at any time slot, up to two queues can be served, each with rate one, but not all three queues can be served simultaneously. Thus, at each time step, the MW policy chooses two queues with the largest backlogs and serves each one of them with rate one.

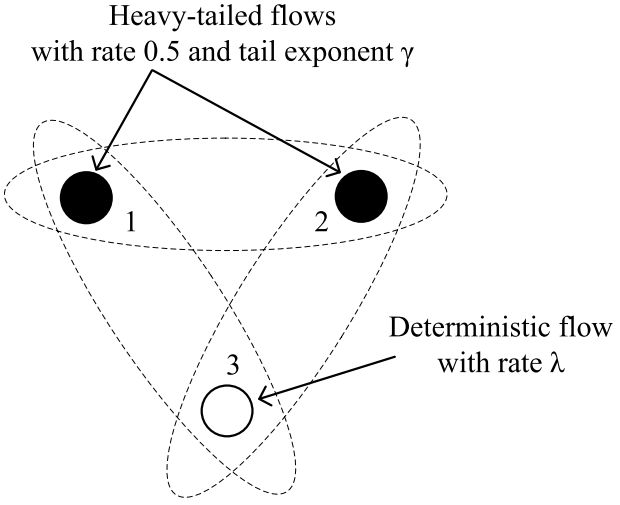
The third queue receives deterministic arrivals with $A_3(t) = \lambda < 1$ for all $t \geq 0$ so that $\gamma_3 = \infty$. The other two queues receive heavy-tailed traffic with a density of the form (4) and tail exponent $\gamma \in (0,1)$ and with rate 0.5; that is, $\mathbb{E}[A_1(t)] = \mathbb{E}[A_2(t)] = 0.5$.

The total arrival rate is $1 + \lambda$, which is less than two, and the network is stable in the conventional sense. The first two queues are automatically delay unstable because they receive heavy-tailed arrivals. However, the third queue can be either delay stable or delay unstable, depending on the values of λ and γ. In what follows, we provide an informal discussion of the different cases.

Case 1 ($0.5 < \lambda < 1$). In this case, queue 3 is delay unstable through a scenario similar to those considered in earlier works (Markakis et al. 2014). Intuitively, because of its heavy-tailed arrivals, Q_1 occasionally receives large inputs. When that happens, with $A_1(t_0)$ being large at some time $t_0 > 0$, Q_1 becomes and stays largest for some time. During that time, the MW policy keeps serving Q_1, while the remaining service capacity is split between Q_2 and Q_3. Because $\lambda > 0.5$, the sum of the arrival rates to Q_2 and Q_3 exceeds the aggregate service rate to these two queues over a time interval of duration $\Omega(A_1(t_0))$. Thus, Q_2 and Q_3 build up to size $\Omega(A_1(t_0))$. Because $\gamma < 1$, we have $\mathbb{E}[A_1^2(t)] = \infty$, and using this property, it can be shown that $\mathbb{E}[Q_3(t)]$ grows unbounded.

The intuition behind this argument is captured by an existing criterion from Markakis et al. (2014) that examines certain trajectories $\mathbf{q}(\cdot)$ of a corresponding fluid model (also called fluid trajectories). In that fluid model, the arrival processes are replaced by deterministic flows with the same rates. The initial conditions are $q_h(0) = 1$ for some heavy-tailed queue (in our example, $h = 1$ or $h = 2$) and $q_j(0) = 0$ for $j \neq h$. (Because of symmetry, we only need to consider the case in which $h = 1$.) Let us say that the "zero fluid" (ZF) condition holds if the solution to the fluid model (known to be unique for the MW policy) keeps $q_3(\cdot)$ at zero. According to the criterion in Markakis et al. (2014), the failure of the ZF condition certifies the delay instability of queue 3. This is indeed the case

Figure 1. A Single-Hop Network with Three Queues, Two of Which Receive Heavy-Tailed Traffic, Whereas the Third One Receives Deterministic Traffic



Notes. In this figure, the dotted ellipses illustrate the different possible service vectors. The third queue is delay unstable for certain ranges of λ and γ.

here: starting with the initial conditions $\mathbf{q}(0) = (1,0,0)$, it can be verified that, for small positive times t, we have $q_3(t) = (\lambda - 0.5)t/2 > 0$.

In summary, for this particular case, we have delay instability, and this is correctly predicted by the failure of the ZF condition and available results.

Case 2 ($0 < \lambda < 0.5$ and $\gamma > 0.5$). In this case, queue 3 turns out to be delay stable (in fact, robustly delay stable). This is a consequence of our general result in Section 5.

For this case, the fluid model initialized at $\mathbf{q}(0) = (1,0,0)$ satisfies $q_3(t) = 0$ for all positive times, and the ZF condition holds. In particular, the ZF condition is "aligned" with delay stability. Observations of this nature lead to the question whether the ZF condition can be used as a certificate of delay stability (Markakis et al. 2018). However, this is not the case as we discuss next.

Case 3 ($0 < \lambda < 0.5$ and $\gamma \le 0.5$). In this case, the ZF condition holds, similar to Case 2. However, queue 3 turns out to be delay unstable; this follows from the proof[2] of our main result, Theorem 1.

We summarize here the underlying intuition. When $\gamma \le 0.5$, there is considerable probability that both queues 1 and 2 receive large inputs within a certain time interval. More concretely, for large values of M, there is probability $\Omega(1/M)$ that both Q_1 and Q_2 receive aggregate arrivals of size at least $3M$ within the time interval $[0, M]$. If this happens, both Q_1 and Q_2 become large enough so that Q_3 receives no service during the interval $[M, 2M]$. As a result, $Q_3(2M) \ge \lambda M$ with probability $\Omega(1/M)$. One can then use this fact to show that, as t increases, $\mathbb{E}[Q_3(t)]$ grows unbounded, and queue 3 is delay unstable.

In summary, the presence or absence of delay stability depends on the tail exponents in a nontrivial manner. Furthermore, the ZF condition cannot discriminate between Cases 2 and 3 and, thus, cannot account for the different outcomes (delay stability in Case 2, delay instability in Case 3). In fact, the same obstacle arises with any other criterion that relies on traditional fluid models because fluid models do not take the tail exponents into account. In order to make progress, we need to consider the probability that large inputs (or jumps) may arrive within a certain time interval as a function of the tail exponents (the probability is larger when the tail exponents are smaller). Furthermore, as illustrated by Case 3, we may have to consider the effect of "coordinated" large jumps at more than one queue within the same time interval. This is accomplished by the model in the next section: it is still in the spirit of traditional fluid models except that it allows for jumps along the heavy-tailed flows, subject to a "budget" on allowed jumps as determined by the tail exponents.

4. Jumping Fluid Models

In this section, we introduce a generalization of the fluid model that allows for jumps along certain coordinates. We proceed by first defining a traditional fluid model and then modifying it.

4.1. The Fluid Model

A fluid model is a deterministic continuous-time dynamical system that replaces the arrival process with a fluid stream of arrivals and updates queue lengths along max-weight drifts. The literature provides a few, somewhat different but equivalent, definitions of the fluid model (Shah and Wischik 2012, Markakis et al. 2018), which typically involve differential equations with boundary conditions. Here, we adopt an equivalent but somewhat simpler definition[3] from Sharifnassab et al. (2020).

Recall the definition of $\mathcal{M}$ as the set of all possible service vectors. For any $\mathbf{x} \in \mathbb{R}_+^\ell$, we define

$$\mathcal{M}(\mathbf{x}) = \{\boldsymbol{\mu} \in \mathcal{M} \,|\, \boldsymbol{\mu}^T\mathbf{x} \ge \boldsymbol{\nu}^T\mathbf{x}, \ \ \forall \boldsymbol{\nu} \in \mathcal{M}\} = \operatorname*{argmax}_{\boldsymbol{\nu}\in\mathcal{M}} \boldsymbol{\nu}^T\mathbf{x}, \tag{6}$$

which is the set of all possible service vectors that, for the given $\mathbf{x}$, attain the maximum in the definition of the MW policy; see (2). We also let

$$\overline{\mathcal{M}}(\mathbf{x}) = \operatorname{conv}(\mathcal{M}(\mathbf{x})). \tag{7}$$

Furthermore, given an arrival rate vector $\boldsymbol{\lambda} \ge \mathbf{0}$ and any $\mathbf{x} \in \mathbb{R}_+^\ell$, we let

$$\overline{\mathcal{D}}_{\boldsymbol{\lambda}}(\mathbf{x}) = \boldsymbol{\lambda} - \overline{\mathcal{M}}(\mathbf{x}) = \operatorname{conv}\left(\left\{\boldsymbol{\lambda} - \boldsymbol{\mu} \,|\, \boldsymbol{\mu} \in \operatorname*{argmax}_{\boldsymbol{\nu}\in\mathcal{M}} \boldsymbol{\nu}^T\mathbf{x}\right\}\right), \tag{8}$$

which is the set of candidate drifts when the queue length vector is $\mathbf{x}$.

For the definitions that follow, recall our convention that $\dot{\mathbf{q}}(t)$ denotes the right derivative of $\mathbf{q}(\cdot)$ with respect to time at time t.

Definition 2 (Fluid Trajectories). Let us fix a network with ℓ nodes, a set $\mathcal{M}$ of possible service vectors, and an arrival rate vector $\boldsymbol{\lambda} \geq \mathbf{0}$. A fluid trajectory corresponding to $\boldsymbol{\lambda}$ is a nonnegative, continuous, and right-differentiable ℓ-dimensional function $\mathbf{q}(\cdot)$ that satisfies the differential inclusion

$$\dot{\mathbf{q}}(t) \in \overline{\mathcal{D}}_{\boldsymbol{\lambda}}(\mathbf{q}(t)), \qquad \forall t \geq 0. \tag{9}$$

Given some $\boldsymbol{\lambda} \geq \mathbf{0}$ and $\mathbf{q}(0) \geq \mathbf{0}$, there always exists a unique (and nonnegative) fluid trajectory $\mathbf{q}(\cdot)$ corresponding to $\boldsymbol{\lambda}$ and initialized at $\mathbf{q}(0)$ (Markakis et al. 2018, Sharifnassab et al. 2020).

4.2. Adding the Jumps

We now introduce JF trajectories. Given a vector $\mathbf{n} = (n_1, \ldots, n_\ell)$ of nonnegative integers, an ϵ-JF($\mathbf{n}$) trajectory is a fluid trajectory with jumps with n_j positive jumps in its jth component, allowing for ϵ-changes in the arrival rate $\boldsymbol{\lambda}$. More concretely, we have the following definition.

Definition 3 (ϵ-Jumping Fluid Trajectories). Let us fix a network with ℓ nodes and a set $\mathcal{M}$ of possible service vectors. We are given an arrival rate vector $\boldsymbol{\lambda}^* \geq \mathbf{0}$, a nonnegative integer vector $\mathbf{n}$, and some $\epsilon \geq 0$. An ϵ-JF($\mathbf{n}$) trajectory corresponding to $\boldsymbol{\lambda}^*$ is a nonnegative ℓ-dimensional function $\mathbf{q}(\cdot)$, which is right-continuous with left limits and right-differentiable initialized with $\mathbf{q}(t) = \mathbf{0}$ for all $t < 0$ and with the following properties:

i. Each component $q_j(\cdot)$ has n_j points of discontinuity.
ii. If $q_j(\cdot)$ has a discontinuity at some time t, then $q_j(t) > q_j(t^-)$.
iii. For every $t \geq 0$,

$$\dot{\mathbf{q}}(t) \in \overline{\mathcal{D}}_{\boldsymbol{\lambda}(t)}(\mathbf{q}(t)), \tag{10}$$

for some nonnegative function $\boldsymbol{\lambda}(\cdot)$ that is right-continuous and piecewise constant with finitely many points of discontinuity and satisfies $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}^*\| \leq \epsilon$ for all t.

Finally, an ϵ-JF($\mathbf{n}$) trajectory with[4] $\boldsymbol{\gamma}^T \mathbf{n} = \sum_{j=1}^{\ell} \gamma_j n_j \leq 1$, is called an ϵ-JF($\boldsymbol{\gamma}$) trajectory.

Note that we allow jumps at time zero, in which case $\mathbf{q}(0) \neq \mathbf{0}$. Note also that, if $\epsilon = 0$ and jumps can only happen at time zero, then an ϵ-JF trajectory is just a fluid trajectory with the jumps of the JF trajectory determining the initial conditions of the fluid trajectory.

More generally, an ϵ-JF trajectory consists of a concatenation of fluid trajectories over the intervals in which $\boldsymbol{\lambda}(\cdot)$ stays constant together with a finite number of jumps. For this reason, once the jump times, jump sizes, and function $\boldsymbol{\lambda}(\cdot)$ are specified, the results for fluid models extend and establish the existence and uniqueness of the ϵ-JF($\mathbf{n}$) trajectory.

Our next definition formalizes the requirement that a certain queue must stay at zero under all ϵ-jumping fluid trajectories.

Definition 4 (JF Conditions). Let us fix a network with ℓ nodes and a set $\mathcal{M}$ of possible service vectors. We are given an arrival rate vector $\boldsymbol{\lambda}^* \geq \mathbf{0}$ and a particular queue, m, of interest.

a. Given a vector $\boldsymbol{\gamma}$ with components in $[0, \infty]$ and some $\epsilon \geq 0$, we say that the ϵ-JF($\boldsymbol{\gamma}$) condition holds for queue m and $\boldsymbol{\lambda}^*$ if, for every ϵ-JF($\boldsymbol{\gamma}$) trajectory corresponding to $\boldsymbol{\lambda}^*$ and every $t \geq 0$, we have $q_m(t) = 0$.

b. Given a vector $\boldsymbol{\gamma}$ with components in $[0, \infty]$, we say that the robust jumping fluid condition (RJF($\boldsymbol{\gamma}$)) holds for queue m and $\boldsymbol{\lambda}^*$ if there exists some $\epsilon > 0$ such that the ϵ-JF($\boldsymbol{\gamma}$) condition holds for queue m and $\boldsymbol{\lambda}^*$.

Note the restriction on the number of jumps in terms of the tail exponents: the heavier the arrival processes (i.e., the smaller the tail exponents γ_j), the larger the number n_j of jumps that we allow. As an example, if $\gamma_1 \leq 1$ and queue 1 is the only heavy-tailed queue, then we allow up to $\lfloor 1/\gamma_1 \rfloor$ jumps at queue 1 and no jumps at the other queues.

5. Main Result

Our main result provides a necessary and sufficient condition for robust delay stability in terms of JF trajectories. The proof is given in Sections 7 and 8.

Theorem 1. *Let us fix a network with ℓ nodes; a set $\mathcal{M}$ of possible service vectors; an arrival rate vector $\boldsymbol{\lambda}^* \geq \mathbf{0}$; a particular queue, m, of interest; and a vector $\boldsymbol{\gamma}$ of tail exponents with components in $(0, \infty]$. The queue m is RDS if and only if the RJF($\boldsymbol{\gamma}$) condition holds for queue m and $\boldsymbol{\lambda}^*$.*

5.1. Some Intuition

We provide here a high-level explanation of our result. Some more refined intuition is provided by the proof outlines in Sections 7.1 and 8.1.

Let M be a large constant. We say that the stochastic process $A_j(t)$ has a jump whenever it is larger than (approximately) M. Let N_j be the number of jumps of $A_j(t)$ during the interval $[0,M]$ and let $\mathbf{N} = (N_1,\ldots,N_\ell)$. It turns out that, for any nonnegative integer vector $\mathbf{n}$, the probability that the event $\mathbf{N} = \mathbf{n}$ scales (approximately) like $M^{-\boldsymbol{\gamma}^T\mathbf{n}}$. The latter quantity is "significant" (in the sense that it makes an unbounded contribution to certain expected values) if and only if $\boldsymbol{\gamma}^T\mathbf{n} \le 1$. Thus, over an interval of length M, we can focus on sample paths for which the realized vector $\mathbf{n}$ of jump counts satisfies $\boldsymbol{\gamma}^T\mathbf{n} \le 1$ and examine whether such sample paths can cause the queue of interest to become large. We then argue that these sample paths are well-approximated by the ϵ-JF($\mathbf{n}$) trajectories involved in the ϵ-JF($\boldsymbol{\gamma}$) condition.

5.2. Remarks

We continue with some remarks on the scope of our result.

5.2.1. Heavy-Tailed Queues. If queue m is heavy-tailed, that is, $\gamma_m \le 1$, then the condition $\boldsymbol{\gamma}^T\mathbf{n} \le 1$ allows $\mathbf{n}$ to be the mth unit vector. With such a vector $\mathbf{n}$, we can have an ϵ-JF($\mathbf{n}$) trajectory with a positive jump in the mth component, resulting in a positive value of $q_m(t)$. Thus, the RJF($\boldsymbol{\gamma}$) condition does not hold, and queue m is not RDS. This is just a variation of the well-known fact that a queue with heavy-tailed arrivals is not delay stable.

5.2.2. Unstable Systems. Theorem 1 makes no stability assumptions. For unstable (or marginally stable) systems, some components of ϵ-JF trajectories can grow arbitrarily large. On the other hand, these components do not necessarily have a substantial effect on the queue, m, of interest. As long as the mth component of all ϵ-JF trajectories stays at zero, queue m is RDS.

5.2.3. Comparison with the ZF Condition. If $\gamma_j > 1/2$ for all j, then an ϵ-JF trajectory can have at most one jump. For stable systems, the RJF condition boils down to a robust version of the ZF condition introduced in Section 3. In other words, for this case, a robust version of the ZF condition is a necessary and sufficient condition for robust delay stability. On the other hand, because the (robust version of) the ZF condition is strictly weaker than the RJF condition, it does not provide necessary and sufficient conditions for general $\boldsymbol{\gamma}$.

5.2.4. The Light-Tailed Case. Suppose that $\gamma_j > 1$ for all j so that all arrival processes are light-tailed. In this case, no jumps are allowed, and the RJF condition boils down to considering ordinary fluid trajectories with slightly perturbed arrival rates. We have the following possibilities:

a. If $\boldsymbol{\lambda}^*$ is in the interior of the capacity region $\overline{\mathcal{M}}$, then $\mathbf{0}$ is in the interior of $\overline{\mathcal{D}}_{\boldsymbol{\lambda}^*}(\mathbf{0}) = \boldsymbol{\lambda}^* - \overline{\mathcal{M}}$. It then turns out that $\mathbf{0}$ is an attracting fixed point of the fluid dynamics, the RJF condition holds, and we have RDS for all queues. This is in line with existing results (e.g., see theorem 4.5 of Georgiadis et al. 2006).

b. If $\boldsymbol{\lambda}^*$ is on the boundary or outside the capacity region and, similar to our earlier discussion of unstable systems, queue m could be either RDS or non-RDS, depending on whether (perturbed) fluid trajectories cause q_m to become positive or not.

5.2.5. The Case of Zero Tail Exponents. Our definitions in Sections 2 and 4 are formulated for a nonnegative vector $\boldsymbol{\gamma}$. However, our result is restricted to the case in which this vector is positive. We comment on the reasons for this.

When $\gamma_j = 0$, we are dealing with an arrival process for which $\mathbb{E}[A_j(t)]$ is finite, whereas $\mathbb{E}[A_j(t)^{1+\gamma}]$ may be infinite for every $\gamma > 0$. Our proofs involve, at certain places, a division by γ_j and break down if $\gamma_j = 0$. It is not clear whether a similar result is possible when some of the tail exponents are zero.

5.2.6. Computational Issues. As already hinted in the introduction, checking the RJF condition algorithmically appears to be a hard computational problem, amenable only to impractical Tarski-like elimination algorithms. In one possible simplification, we might just consider ϵ-JF trajectories with $\epsilon = 0$ so that $\boldsymbol{\lambda}(t) = \boldsymbol{\lambda}^*$ for all times t. However, this would still leave the indeterminacy of the jump times and sizes to be reckoned with. Even worse, a restriction to this limited class of trajectories does not seem to lead to necessary and sufficient conditions for any suitably modified notion of stability. See Appendix A.2 for further discussion. A related question is whether we could, without loss of generality, require all the jumps to occur at the same time, for example, at time zero, thus eliminating the need to consider all possible values of the jump times. Unfortunately, this is not the

case: there exist examples in which ϵ-JF trajectories can drive a queue m to a positive value, but this can happen only if we allow the jumps to occur at different times; see Appendix A.5 for an example.

5.3. Our Example, Revisited

On the positive side, Theorem 1 elucidates the precise mechanism that leads to delay instability through a coordination of multiple abnormally large arrival vectors at possibly different times and queues. Furthermore, it allows us to analyze simple problems, such as the one discussed in Section 3, which we do next.

Recall the network of three queues in Section 3, in which $\boldsymbol{\gamma} = (\gamma, \gamma, \infty)$. For γ in the range $(0, 0.5]$, consider a jumping fluid trajectory $\mathbf{q}(\cdot)$ with $\mathbf{n} = (1, 1, 0)$ in which both q_1 and q_2 undergo unit jumps at time 0. With this trajectory, q_3 immediately starts to grow positive. Because $\boldsymbol{\gamma}^T \mathbf{n} = 2\gamma \le 1$, it follows that the RJF($\boldsymbol{\gamma}$) condition fails to hold. Theorem 1 then establishes that queue 3 is not robustly delay stable.

On the other hand, when γ is in the range $(0.5, 1]$, the constraint $\boldsymbol{\gamma}^T \mathbf{n} \le 1$ allows at most one jump in either q_1 or q_2. Without loss of generality, we can assume that this jump takes place at time 0. It turns out that the fluid trajectories that start from either $\mathbf{q}(0) = (1, 0, 0)$ or $\mathbf{q}(0) = (0, 1, 0)$ keep $q_3(\cdot)$ at zero if and only if $\lambda \le 0.5$. Therefore, for γ in range $(0.5, 1]$ and also taking robustness into account, queue 3 is RDS if and only if $\lambda < 0.5$.

6. Robust Delay Stability via Lyapunov Functions

Lyapunov functions are a powerful tool for the stability analysis of queueing networks (Tassiulas and Ephremides 1992, Bertsimas et al. 2001, Maguluri et al. 2016), for example, in throughput optimality proofs for the MW policy (Tassiulas and Ephremides 1992, Neely 2010). Markakis (2013) and Markakis et al. (2018) provide a sufficient condition for delay stability based on a class of piecewise linear Lyapunov functions and use it to derive a sharp characterization of delay stability for a special class of networks, namely, networks with disjoint schedules. Nevertheless, the Lyapunov approach in Markakis (2013) and Markakis et al. (2018) has some drawbacks: (a) the condition provided therein is, in general, only sufficient for delay stability; (b) it does not take into account the tail exponents even though they play an essential role in delay stability as already discussed in Sections 3 and 5.3; (c) the Lyapunov functions considered are piecewise linear, which is perhaps inadequate for the purpose of tight delay stability conditions.

In this section, we explore the power of Lyapunov functions for the case in which the tail exponents of the heavy-tailed queues may be arbitrarily close to zero so that the RJF condition allows an arbitrarily large number of jumps.

For the remainder of this section, we assume that queues $1, \dots, h$ can be heavy-tailed, where $h < \ell$, whereas the remaining queues are light-tailed. Formally, we consider the set Γ of tail coefficients, defined by

$$\Gamma = \{\boldsymbol{\gamma} > \mathbf{0} : \gamma_j > 1, \text{ for } j = h+1, \dots, \ell\}.$$

We also fix an arrival rate vector $\boldsymbol{\lambda}^* \ge \mathbf{0}$ and a light-tailed queue $m > h$ of interest.

Definition 5 (Special Lyapunov Function). For any $\epsilon > 0$, we say that a function $V : \mathbb{R}_+^\ell \to \mathbb{R}_+$ is a special ϵ-Lyapunov function if

1. V is Lipschitz continuous with Lipschitz constant one.
2. $\dot{V}(\mathbf{q}(t)) \le -\epsilon$ whenever $V(\mathbf{q}(t)) > 0$ for all fluid trajectories $\mathbf{q}(\cdot)$ corresponding to arrival rate $\boldsymbol{\lambda}^*$.
3. We have $V(\mathbf{0}) = 0$. Furthermore, if $q_m > 0$, then $V(q_m) > 0$.
4. V is nonincreasing along the coordinates associated with heavy-tailed queues; that is, for $j = 1, \dots, h$, for any $\mathbf{q} \in \mathbb{R}_+^n$, and any $\alpha > 0$, we have $V(\mathbf{q} + \alpha \boldsymbol{e}_j) \le V(\mathbf{q})$, where $\boldsymbol{e}_j$ is the jth unit vector.

Special ϵ-Lyapunov functions as defined are quite similar to the functions considered in theorem 2 of Markakis et al. (2018). However, in contrast to Markakis et al. (2018), our special Lyapunov functions need not be piecewise linear. Our next result establishes a strong connection between the ϵ-JF($\boldsymbol{\gamma}$) condition and the existence of special ϵ-Lyapunov functions. The proof is given in Section 9.

Theorem 2. *For any $\epsilon > 0$, there exists a special ϵ-Lyapunov function if and only if the ϵ-JF($\boldsymbol{\gamma}$) condition holds for every $\boldsymbol{\gamma} \in \Gamma$.*

The special ϵ-Lyapunov functions constructed in the proof of Theorem 2 are not piecewise linear. We do not know whether Theorem 2 remains valid if we restrict to piecewise linear functions.

Combining Theorems 1 and 2, we can establish a strong connection between robust delay stability and special ϵ-Lyapunov functions. In what follows, we say that queue m is ϵ-RDS($\boldsymbol{\gamma}$) if it is delay stable under all arrival processes in the class $\mathcal{A}_\epsilon(\boldsymbol{\gamma}; \boldsymbol{\lambda}^*)$ in Definition 1(a).

Corollary 1. *Let us fix a network with ℓ nodes, a set $\mathcal{M}$ of possible service vectors, an arrival vector $\boldsymbol{\lambda}^* \geq \mathbf{0}$, the number h of heavy-tailed queues, and a light-tailed queue $m > h$.*

a. *If there exists a special ϵ-Lyapunov function for some $\epsilon > 0$, then queue m is RDS for all tail exponents $\boldsymbol{\gamma} \in \Gamma$.*

b. *Suppose that there exists some $\epsilon > 0$ such that, for all $\boldsymbol{\gamma} \in \Gamma$, queue m is ϵ-RDS($\boldsymbol{\gamma}$). Then, there exists an ϵ-special Lyapunov function.*

The proof is provided in Section 9.3. We conjecture that Corollary 1(b) can be strengthened to provide a converse to part (a).

Hypothesis 1. *If queue m is RDS for all tail exponents $\boldsymbol{\gamma} \in \Gamma$, then for some $\epsilon > 0$, there exists a special ϵ-Lyapunov function.*

If Hypothesis 1 is true, we have a tight characterization: a queue is RDS for all tail exponents $\boldsymbol{\gamma} \in \Gamma$ if and only if there exists a special ϵ-Lyapunov function that testifies to this. Establishing the conjecture appears to be difficult. Technically it amounts to reversing the order of the quantifiers in the clause "there exists $\epsilon > 0$ such that, for all $\boldsymbol{\gamma} \in \Gamma \ldots$" in Corollary 1(b) and showing equivalence with the statement "for every $\boldsymbol{\gamma} \in \Gamma$, there exists some $\epsilon > 0 \ldots$"

Our Lyapunov-based results are relevant to the case in which nothing is known about the tail exponents of the heavy-tailed queues other than the fact that they are positive. On the other hand, Lyapunov functions are unlikely to provide useful characterizations of robust delay stability for specific values of the tail exponents because there is no apparent way of accounting for the number of jumps through Lyapunov functions.

7. Proof of the "if" Direction of Theorem 1 (RJF⇒RDS)

In this section, we provide the proof of the "if" direction of Theorem 1, that is, that the RJF condition implies RDS.

Throughout this section, we consider a network with ℓ nodes and a set $\mathcal{M}$ of possible service vectors. We fix an arrival rate vector $\boldsymbol{\lambda}^* \geq \mathbf{0}$; a particular queue, m, of interest; a vector $\boldsymbol{\gamma}$ of tail exponents with components in $(0, \infty]$; and some $\epsilon > 0$ for which the ϵ-JF($\boldsymbol{\gamma}$) condition holds for $\boldsymbol{\lambda}^*$ and queue m. Our goal is to establish robust delay stability for queue m.

The proof is organized in a sequence of lemmas whose proofs are collected in Appendix B. However, before proceeding to the formal arguments, it is helpful to provide an overview of the proof.

7.1. Outline of the Proof

Let us fix an arbitrary time T. We aim at upper bounds for the probability $\mathbb{P}(Q_m(T) \geq M)$ as M gets large. As long as these bounds are a summable function of M and independent of T it follows that $\mathbb{E}[Q_m(T)]$ is finite and a bounded function of T, which is our goal. Let us now fix some M and keep it fixed throughout except for the end of the proof.

The proof relies on various probabilistic bounds as well as on deterministic properties of the MW dynamics. Let us start with the probabilistic part, which is focused on showing that the stochastic system mostly follows the deterministic fluid dynamics except for certain jumps caused by the heavy tails of the arrival processes. We define a threshold for what constitutes a jump and then develop a probabilistic bound on the numbers of jumps. A difficulty here is that, if we use a fixed threshold and because T is arbitrary, a bound on the number of jumps is not possible. We handle this issue by using a threshold θ_t that increases almost linearly as we move further to the past of the form

$$\theta_t = \frac{M + T - t}{\eta \log (M + T - t)}, \qquad t = 0, 1, \ldots, T. \tag{11}$$

This is for some positive constant η to be defined later. At any time $t \leq T$ and for any index j, we say that $A_j(t)$ is a jump if $A_j(t) > \theta_t$. Ignoring logarithmic factors, we show that the jump probability $\mathbb{P}(A_j(t) > \theta_t)$ is of order at most $1/(M + T - t)^{1+\gamma'_j}$, where γ'_j is slightly smaller than the tail exponent γ_j. By summing over t and after some elementary calculations, we then obtain that $\mathbb{P}(N_j = n_j)$ is of order at most $1/M^{n_j \gamma'_j}$, where N_j is the number of jumps of the jth arrival process during the interval $[0, T]$. Then, a further calculation shows that, if $\mathbf{N} = (N_1, \ldots, N_\ell)$, then $\mathbb{P}(\boldsymbol{\gamma}^T \mathbf{N} > 1)$ is of order at most $1/M^\beta$ for some constant $\beta \in (1, 2)$ and is, therefore, a summable function of M; see Lemma 1.

We then consider stochastic fluctuations in the arrival process other than jumps. We argue that they average out so that the cumulative arrival process follows its fluid counterpart. We refer to this as the "small fluctuations event," and show that it occurs with probability at least $1-\ell/M^2$; see (21) and Lemma 3.

Having completed the probabilistic analysis, we then switch to deterministic (sample path) considerations. Let $W(\mathbf{n})$ be the set of all points $\mathbf{q}$ that can be reached by some ϵ-JF($\mathbf{n}$) trajectory (see Definition 6). As a first step, we exploit some special properties of the MW dynamics and show that $W(\mathbf{n})$ is ϵ-attracting; that is, any fluid trajectory that starts outside $W(\mathbf{n})$ moves toward that set with rate at least ϵ (Lemma 4). We then consider a "nice" sample path, that is, a sample path for which the small fluctuations event occurs and for which the realized vector of jump counts $\mathbf{n}$ satisfies $\boldsymbol{\gamma}^T\mathbf{n}\le 1$. (As discussed earlier, nice sample paths have probability $1-O(M^{-\beta})$ with $\beta>1$.) Our deterministic analysis, outlined in the next paragraph, shows that any such sample path stays within $O(M)$ distance from the set $W(\mathbf{n})$. Because $\boldsymbol{\gamma}^T\mathbf{n}\le 1$, the ϵ-JF($\boldsymbol{\gamma}$) condition implies that any point in $W(\mathbf{n})$ satisfies $q_m=0$. Thus, for nice sample paths, we have $Q_m(T)=O(M)$, and therefore, $\mathbb{P}(Q_m(T)\ge M)$ is of order at most $1/M^{\beta}$ for the constant $\beta\in(1,2)$ mentioned earlier. This readily implies a uniform upper bound on $\mathbb{E}[Q_m(T)]$.

The analysis of the dynamics under nice sample paths has two parts. We first study the dynamics between jumps: we rely on the small fluctuations event and then make use of a result from Sharifnassab et al. (2020, theorem 4) to ensure that small fluctuations in the arrivals result into comparably small changes in the resulting stochastic trajectory; cf. Lemma 5. Second, to understand what happens at jump times, we recall that the jump vectors $\mathbf{n}$ associated to nice sample paths satisfy $\boldsymbol{\gamma}^T\mathbf{n}\le 1$. Such vectors $\mathbf{n}$ are allowed in ϵ-JF($\boldsymbol{\gamma}$) trajectories, and therefore, the jumps cannot take the ϵ-JF($\mathbf{n}$) trajectory away from $W(\mathbf{n})$. This implies that a nice sample path stays "close" to an ϵ-JF($\mathbf{n}$) trajectory and, therefore, has a "small" Q_m.

7.2. Sensitivity of Max-Weight Dynamics

The proof for both directions of the theorem requires fairly precise bounding of the fluctuations of the stochastic trajectories. To this effect, we rely heavily on a fluctuation bound for the MW dynamics established in Sharifnassab et al. (2020).

Theorem 3 (Sharifnassab et al. 2020, theorem 2). *Fix a network (i.e., the number of nodes and the set $\mathcal{M}$ of possible service vectors) operating under the MW policy and let $\mathbf{Q}(\cdot)$ be the corresponding queue length stochastic process. There exists a (deterministic) constant $C\ge 1$ such that, for any arrival rate vector $\boldsymbol{\lambda}\ge\mathbf{0}$, any $\mathbf{q}(0)\ge\mathbf{0}$, any $t\ge 0$, and any sample path, if $\mathbf{q}(0)=\mathbf{Q}(0)$, then*

$$\|\mathbf{Q}(t)-\mathbf{q}(t)\|\le C\left(1+\|\boldsymbol{\lambda}\|+\max_{k<t}\left\|\sum_{\tau=0}^{k}(\mathbf{A}(\tau)-\boldsymbol{\lambda})\right\|\right), \tag{12}$$

where $\mathbf{q}(\cdot)$ is the fluid trajectory corresponding to $\boldsymbol{\lambda}$ initialized at $\mathbf{q}(0)$.

We actually use the following variant of Theorem 3, which allows for different initial conditions. The proof is given in Appendix B.1.

Theorem 4. *Under the same assumptions as in Theorem 3 except that we allow for $\mathbf{Q}(0)$ and $\mathbf{q}(0)$ to be different and for the same constant C, we have*

$$\|\mathbf{Q}(t)-\mathbf{q}(t)\|\le\|\mathbf{Q}(0)-\mathbf{q}(0)\|+C\left(1+\|\boldsymbol{\lambda}\|+\max_{k<t}\left\|\sum_{\tau=0}^{k}(\mathbf{A}(\tau)-\boldsymbol{\lambda})\right\|\right).$$

7.3. Arrival Process and Jumps

We now return to the formal proof. Recall that, throughout this section, we fix the network, $\boldsymbol{\lambda}^*\ge\mathbf{0}$ and the tail exponents $\gamma_j\in(0,\infty]$. We assume that the ϵ-JF($\boldsymbol{\gamma}$) condition holds for $\boldsymbol{\lambda}^*$, queue m, and some $\epsilon>0$. We fix some positive integers M and T; these remain fixed throughout except for the end of the proof and for some additional assumptions that M is "large enough."

We consider an arrival process $\mathbf{A}(\cdot)$ in the class $\mathcal{A}_\delta(\boldsymbol{\gamma};\boldsymbol{\lambda}^*)$ introduced in Definition 1 with $\delta=\gamma\epsilon/20C$, where C is the constant in Theorem 4 and

$$\gamma=\min_{j=1,\dots,\ell}\gamma_j. \tag{13}$$

In particular,

$$\|\mathbb{E}[\mathbf{A}(t)] - \boldsymbol{\lambda}^*\| \le \frac{\gamma\epsilon}{20C}, \qquad \forall t \ge 0. \tag{14}$$

Our goal is to derive an upper bound on $\mathbb{P}(\mathbf{Q}_m(T) \ge M)$ that holds uniformly for every arrival process $\mathbf{A}(\cdot)$ in $\mathcal{A}_\delta(\boldsymbol{\gamma}; \boldsymbol{\lambda}^*)$, every T, and every large enough M. This is then used to conclude that the RDS property holds.

Because we assume that the tail exponents are nonzero, we have $\gamma > 0$. Furthermore, in order to simplify the proof, it is convenient to assume that

$$\gamma \le 1. \tag{15}$$

Claim 1. *The assumption $\gamma \le 1$ can be made without loss of generality.*

Proof. Given a system, call it S, consider a new system S' in which we add one more queue, queue 0, with $\gamma_0 < 1$, which does not interact with the others; for example, for any allowed service vector $\boldsymbol{\mu}$ in system S, we introduce a corresponding service vector $(1, \boldsymbol{\mu})$ in system S'. This way, the dynamics of queues $1, \dots, \ell$ are the same in the two systems S and S'. It follows that, for $m \ge 1$, queue m is RDS in system S if and only if it is RDS in system S'. Furthermore, because of the lack of interaction, the RJF condition for queue m holds in system S if and only if it holds in system S'.

Once we prove the result for the case $\gamma \le 1$, we apply it to system S' and obtain the equivalence of RDS and RJF for the latter system. Based on this discussion, this also establishes the equivalence of RDS and RJF for the original system S. □

We define some more constants:

$$\overline{\mu} = 1 + \max_{\boldsymbol{\mu} \in \mathcal{M}} \|\boldsymbol{\mu}\| + \|\boldsymbol{\lambda}^*\| + \epsilon, \tag{16}$$

and

$$\eta = \frac{8000\, C^2 \ell^2 \overline{\mu}}{(\gamma\epsilon)^2}. \tag{17}$$

Having fixed M, T, and η, we finally define θ_t as in (11).

For $j = 1, \dots, \ell$ and $t = 0, \dots, T-1$, we say that t is a jump time for $A_j(\cdot)$ if $A_j(t) > \theta_t$. For any j and any $\tau \in [0, T)$, we let $N_j(\tau)$ be the (random) number of jumps in $A_j(\cdot)$ that occur during $[0, \tau]$ and also define the corresponding vector $\mathbf{N}(\tau) = (N_1(\tau), \dots, N_\ell(\tau))$. To simplify notation and as long as T is fixed, we use $\mathbf{N}$ and N_j to refer to $\mathbf{N}(T-1)$ and $N_j(T-1)$, respectively. Note that, with this definition, $\mathbf{N} = \mathbf{N}(T-1)$ includes all jumps that can affect $\mathbf{Q}(T)$.

We consider the event

$$\mathcal{E}^{\text{jump}}(T, M) = \{\boldsymbol{\gamma}^T \mathbf{N} \le 1\}. \tag{18}$$

We argue that it occurs with high probability.

Let F be the set of all $\boldsymbol{n} \ge \mathbf{0}$ such that $1 < \boldsymbol{\gamma}^T \boldsymbol{n} \le 2$. If F is empty, then whenever $\boldsymbol{\gamma}^T \boldsymbol{n} > 1$, we must have $\boldsymbol{\gamma}^T \boldsymbol{n} > 2$. If F is nonempty, it has finitely many elements, which implies that the minimum $\min_{\boldsymbol{n} \in F} \boldsymbol{\gamma}^T \boldsymbol{n}$ is attained and its value is greater than one. In either case, we obtain that there exists a constant $\beta > 1$ such that

$$\text{if } \boldsymbol{\gamma}^T \boldsymbol{n} > 1, \text{ then } \boldsymbol{\gamma}^T \boldsymbol{n} \ge \beta^3. \tag{19}$$

Without loss of generality, we can take β to satisfy $1 < \beta < 2$. In the next lemma, we show the probability that $\boldsymbol{\gamma}^T \mathbf{N} > 1$ decays at least as fast as $1/M^\beta$.

Lemma 1. *There exists a constant $M_1 \ge 0$ independent of T such that, if $M \ge M_1$, then*

$$\mathbb{P}(\mathcal{E}^{\text{jump}}(T, M)) = \mathbb{P}(\boldsymbol{\gamma}^T \mathbf{N} \le 1) \ge 1 - M^{-\beta}.$$

The proof is given in Appendix B.2 and relies on the intuition that the probability of a jump in the jth arrival process scales at most as $M^{-\gamma_j}$ (approximately), which then implies that the probability of the event $\{\mathbf{N} = \mathbf{n}\}$ scales at most as $M^{-\boldsymbol{\gamma}^T \mathbf{n}}$ (approximately).

7.4. Fluctuations of the Arrival Process

We now study the remaining fluctuations of the cumulative arrival processes after we exclude the jumps. We begin with a concentration inequality for the sum of independent random variables. The proof of our next lemma

is given in Appendix B.3 and is essentially a reformulation of the Bernstein inequality; see, for example, (1.21) in appendix 1 of Anthony and Bartlett (2009).

Lemma 2. *Suppose that $X_1, \ldots, X_n$ are independent random variables that satisfy $X_i \in [0, b]$ and $\mathbb{E}[X_i] \le \overline{\lambda}$ for some $b, \overline{\lambda} > 0$. Let $Y = X_1 + \cdots + X_n$. Then, for any $z \ge 0$,*

$$\mathbb{P}(|Y - \mathbb{E}[Y]| > z) \le 2\exp\left(-\frac{z^2}{2b(\overline{\lambda}n + z/3)}\right). \tag{20}$$

We consider the truncated process $\mathbf{A}^*(t) = \min\{\mathbf{A}(t), \theta_t\}$, where the minimum is taken component-wise. We define the small fluctuations event

$$\mathcal{E}^{\text{fluc}}(T, M) = \left\{ \left\| \sum_{\tau=t_0}^{t} (\mathbf{A}^*(\tau) - \boldsymbol{\lambda}^*) \right\| \le \frac{\gamma\epsilon}{10C}(M + T - t_0), \text{ for } 0 \le t_0 \le t < T \right\}, \tag{21}$$

where C is the constant in Theorem 4.

Lemma 3. *There exists a constant $M_2 \ge 0$ independent of T such that, if $M \ge M_2$, then $\mathbb{P}(\mathcal{E}^{\text{fluc}}(T, M)) \ge 1 - \ell M^{-2}$.*

The proof is somewhat long but straightforward. It relies on the concentration inequality in Lemma 2 and is given in Appendix B.4.

7.5. Deterministic Analysis of the Dynamics

We start with some definitions. It is useful to recall here that ϵ-JF trajectories start at zero just before time 0 but can become nonzero at time 0 or later because of jumps or unstable drifts.

Definition 6. For any nonnegative integer vector $\mathbf{n}$, we define $W(\mathbf{n})$ as the set of all points in $\mathbb{R}_+^\ell$ that can be reached by some ϵ-JF($\mathbf{n}$) trajectory.

Because ϵ-JF trajectories are by definition nonnegative, $W(\mathbf{n})$ is a subset of $\mathbb{R}_+^\ell$. Note that the set $W(\mathbf{n})$ depends on ϵ, but because ϵ is held fixed, we suppress this dependence from our notation.

Definition 7 (ϵ-Attracting and ϵ-Invariant Sets).

a. A subset W of $\mathbb{R}_+^\ell$ is ϵ-attracting if every fluid trajectory $\mathbf{q}(\cdot)$ corresponding to $\boldsymbol{\lambda}^*$ and initialized at some arbitrary $\mathbf{q}(0) \ge \mathbf{0}$ satisfies

$$\frac{d}{dt} d(\mathbf{q}(t), W) \le -\epsilon, \tag{22}$$

whenever $\mathbf{q}(t) \notin W$.

b. A subset W of $\mathbb{R}_+^\ell$ is ϵ-invariant if, for any $\boldsymbol{\lambda} \ge \mathbf{0}$ that satisfies $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\| \le \epsilon$ and any fluid trajectory $\mathbf{q}(\cdot)$ corresponding to $\boldsymbol{\lambda}$ and initialized at some arbitrary $\mathbf{q}(0) \in W$, we have $\mathbf{q}(\tau) \in W$ for all $\tau \ge 0$.

We observe that if W is ϵ-attracting and $0 \le t_0 < t_1$, then every fluid trajectory satisfies

$$d(\mathbf{q}(t_1), W) \le \max\{0, d(\mathbf{q}(t_0), W) - (t_1 - t_0)\epsilon\}. \tag{23}$$

It can be shown that an ϵ-attracting set is ϵ-invariant, but we do not need this fact. We are interested instead in the converse statement, that every ϵ-invariant set is ϵ-attracting, which we then apply to the set $W(\mathbf{n})$; this is the subject of the next lemma.

Lemma 4.

i. *Every ϵ-invariant set is ϵ-attracting.*

ii. *For any $\mathbf{n} \in \mathbb{Z}_+^\ell$, the set $W(\mathbf{n})$ in Definition 6 is ϵ-invariant and, a fortiori, ϵ-attracting.*

The proof is given in Appendix B.5 and relies on the intuition that, for an ϵ-invariant set W, ϵ-perturbations of $\boldsymbol{\lambda}^*$ cannot take the trajectories away from W. This means that there must be a drift toward W that locally counteracts such perturbations. The nonexpansive property of the max-weight dynamics then enables us to extend this local counteraction result into the desired global attraction property.

Let us fix a sample path of the arrival process. For any $t \in [0, T)$, let $\mathbf{n}(t)$ be the realized value of $\mathbf{N}(t)$, that is, $\mathbf{n}(t)$ is the vector with the realized number of jumps in the process $\mathbf{A}(\cdot)$ up to time t (see the paragraph preceding (18)). With a bit of abuse of notation, we let

$$W(t) = W(\mathbf{n}(t-1)). \tag{24}$$

(The reason for the $t - 1$ term on the right-hand side is that we want to compare $W(\cdot)$ and $\mathbf{Q}(\cdot)$, but $\mathbf{Q}(t)$ is only affected by jumps that happen before time t.)

7.6. Some More Intuition

The general idea is to show that $\mathbf{Q}(\cdot)$ stays close to the set $W(\cdot)$ so that we can ultimately exploit the fact that $q_m = 0$ for every $\mathbf{q} \in W(T)$. There are two parts to the argument:

i. If, at a certain time, $Q_j(t)$ has a jump (i.e., a large increase), the set $W(t)$ expands along the jth coordinate, and so the distance between $\mathbf{Q}(t)$ and $W(t)$ does not increase.

ii. In between jumps, $\mathbf{Q}(t)$ follows the fluid trajectory, plus some fluctuations, within the range allowed by Lemma 3. These fluctuations get "eliminated" because the fluid trajectory is attracted to $W(t)$ (Lemma 4). Note that (21) allows for larger fluctuations in the far past (see the term $M + T - t_0$); however, for fluctuations in the far past, the motion in the direction of $W(t)$ happens for a longer time period, enough to eliminate them. This explains the choice of the threshold θ_t in (11); the logarithmic term in the denominator is included for technical reasons.

7.7. The Distance from the Invariant Set

Given times that satisfy $0 \le t_0 < t_1 \le T$, we say that the interval (t_0, t_1) is jump-free if $A_j(\tau) \le \theta_\tau$ for all j and all $\tau \in (t_0, t_1)$. Note that the initial time t_0 and the end time t_1 are allowed to be jump times. Let $M_3 \ge 1$ be a constant, independent of T, such that, for any $M' \ge M_3$,

$$\frac{\gamma\epsilon M'}{10} + \frac{\ell M'}{\eta \log M'} + \epsilon + (\|\boldsymbol{\lambda}^*\| + 1)C + \overline{\mu} \le \frac{\gamma\epsilon M'}{6}. \tag{25}$$

Lemma 5. *Suppose that $M \ge M_3$ and fix times that satisfy $0 \le t_0 < t_1 \le T$. Consider a sample path under which the event $\mathcal{E}^{\text{fluc}}(T, M)$ occurs and the interval (t_0, t_1) is jump-free. Then,*

$$d(\mathbf{Q}(t_1), W(t_1)) \le \max\{0, d(\mathbf{Q}(t_0), W(t_0)) - (t_1 - t_0)\epsilon\} + \frac{\epsilon\gamma}{6}(M + T - t_0). \tag{26}$$

Note that the lemma also applies when $t_1 = t_0 + 1$ so that the set of integers in (t_0, t_1) is empty. The proof, which is given in Appendix B.6, relies on the sensitivity bound in Theorem 4 and the fact that $W(t_0)$ is ϵ-attractive. The first term on the right-hand side reflects the fact that a fluid trajectory is attracted to $W(\cdot)$ during the jump-free interval; the second term reflects the effect of the smaller fluctuations during the interval (t_0, t_1).

We then apply Lemma 5 and use strong induction on t for $t \le T$ to establish the next lemma. Its proof is given in Appendix B.7.

Lemma 6. *Suppose that $M \ge M_3$. Consider a sample path under which the events $\mathcal{E}^{\text{jump}}(T, M)$ and $\mathcal{E}^{\text{fluc}}(T, M)$ occur. Then, $d(\mathbf{Q}(T), W(T)) \le M\epsilon/2$.*

7.8. Bounding $\mathbb{E}[\mathbf{Q}_m(t)]$

Let $\overline{M} = \max\{M_1, M_2, M_3\}$, where M_1, M_2, and M_3 are the constants in Lemma 1, Lemma 3, and (25), respectively. Because these constants are independent of T, the constant $\overline{M}$ is also independent of T.

Let us consider some $M \ge \overline{M}$ and a sample path under which the events $\mathcal{E}^{\text{jump}}(T, M)$ and $\mathcal{E}^{\text{fluc}}(T, M)$ occur. In particular, we have $\boldsymbol{\gamma}^T\mathbf{n} \le 1$, where $\mathbf{n}$ is the realized value of $\mathbf{N}$, that is, the vector with the number of jumps until time $T - 1$ for that particular sample path. The ϵ-JF($\boldsymbol{\gamma}$) condition, which we assume to hold, implies that every ϵ-JF($\mathbf{n}$) trajectory with $\boldsymbol{\gamma}^T\mathbf{n} \le 1$ keeps q_m at zero, and therefore, every vector $\mathbf{q} \in W(T)$ has $q_m = 0$. Consequently, for every sample path in $\mathcal{E}^{\text{jump}}(T, M) \cap \mathcal{E}^{\text{fluc}}(T, M)$, we have

$$Q_m(T) \le d(\mathbf{Q}(T), W(T)) \le \frac{M\epsilon}{2}, \tag{27}$$

where the second inequality follows from Lemma 6. Therefore, for any $M \ge \overline{M}$, we have

$$\begin{aligned}\mathbb{P}\left(Q_m(T) > \frac{M\epsilon}{2}\right) &\le (1 - \mathbb{P}(\mathcal{E}^{\text{jump}}(T, M))) + (1 - \mathbb{P}(\mathcal{E}^{\text{fluc}}(T, M))) \\ &\le M^{-\beta} + \ell M^{-2},\end{aligned} \tag{28}$$

where the first inequality follows from (27) and the union bound and the second inequality is due to Lemmas 1 and 3. Because β is by definition greater than one, the formula $\mathbb{E}[Q_m(T)] = \int_0^\infty \mathbb{P}(Q_m(t) > M)\, dM$ implies that $\mathbb{E}[Q_m(T)]$ is bounded above by a constant that does not depend on T. Furthermore, note that this bound applies uniformly to all processes in the class $\mathcal{A}_\delta(\boldsymbol{\gamma}; \boldsymbol{\lambda}^*)$ for $\delta = \gamma\epsilon/20C$ (see (14)). This shows that queue m is robustly delay stable and completes the proof of the first direction of Theorem 1.

8. Proof of the Reverse Direction of Theorem 1 (RDS ⇒ RJF)

In this section, we prove the reverse ("only if") direction of Theorem 1, that is, that RDS implies that the ϵ-JF($\boldsymbol{\gamma}$) condition holds for some $\epsilon > 0$. The proof is organized in a sequence of lemmas whose proofs are collected in Appendix C.

We actually prove the contrapositive and start by assuming that, for every $\epsilon > 0$, there exists an ϵ-JF($\mathbf{n}$) trajectory $\mathbf{q}^\epsilon(\cdot)$ with $\boldsymbol{\gamma}^T\mathbf{n} \le 1$ and such that $\mathbf{q}^\epsilon_m(t) > 0$ at some positive time t.

We keep $\boldsymbol{\lambda}^*$, $\boldsymbol{\gamma}$, and ϵ fixed throughout the proof and show that there exists an arrival processes in the class $\mathcal{A}_\epsilon(\boldsymbol{\gamma}; \boldsymbol{\lambda}^*)$ for which queue m is not delay stable. Because ϵ can be arbitrarily small, this implies that there exists no δ such that queue m is delay stable for all arrival processes in the class $\mathcal{A}_\delta(\boldsymbol{\gamma}; \boldsymbol{\lambda}^*)$, and therefore, queue m is not RDS. However, before proceeding to the formal arguments, we overview informally the key ideas in the proof.

8.1. Outline of the Proof

The main idea is to construct a certain arrival process $\mathbf{A}(\cdot)$ in the class $\mathcal{A}_\epsilon(\boldsymbol{\gamma}; \boldsymbol{\lambda}^*)$, whose arrival rate is a time-scaled version of the piecewise constant rate $\boldsymbol{\lambda}(\cdot)$ associated with the ϵ-JF($\mathbf{n}$) trajectory. We then use the bounded sensitivity property of the MW dynamics (Theorem 4) to show that the resulting process $\mathbf{Q}(\cdot)$ tracks a suitably scaled (by a factor of T) version of the ϵ-JF trajectory $\mathbf{q}^\epsilon(\cdot)$ with substantial probability. In particular, $\mathbf{Q}(T)$ is (with substantial probability) comparable to $T\mathbf{q}^\epsilon(1)$, leading to a large value of $\mathbb{E}[Q_m(T)]$. This part of the argument capitalizes on the fact that the number of jumps of the ϵ-JF($\mathbf{n}$) trajectory is limited by the condition $\boldsymbol{\gamma}^T\mathbf{n} \le 1$.

On the technical side, the tracking result involves two separate arguments:

a. Whenever the ϵ-JF($\mathbf{n}$) trajectory has a jump, at some time τ, there is substantial probability that the stochastic process also has a jump at some time near the scaled counterpart, $T\tau$, of τ; see Lemma 8.

b. In between jump times of the ϵ-JF($\mathbf{n}$) trajectory, we use concentration inequalities to show that there is a fairly large probability that the stochastic process stays close to its fluid counterpart.

8.2. Jumps of the ϵ-JF(n) Trajectory

We fix a network with ℓ nodes; a set $\mathcal{M}$ of possible service vectors; an arrival rate vector $\boldsymbol{\lambda}^* \ge \mathbf{0}$; a particular queue, m, of interest; and a vector $\boldsymbol{\gamma}$ of tail exponents with components in $(0, \infty]$. We also fix some $\epsilon > 0$ and assume that the ϵ-JF($\boldsymbol{\gamma}$) condition fails to hold.

We start with a few elementary observations, namely, that the ϵ-JF($\mathbf{n}$) trajectories of interest can be taken, without loss of generality, through scaling and perturbations, to have some convenient properties that allow us to simplify subsequent notation and arguments. The proof is given in Appendix C.1.

Lemma 7. *If the ϵ-JF($\boldsymbol{\gamma}$) condition fails to hold, then there exists some $\mathbf{n} \ge \mathbf{0}$ with $\boldsymbol{\gamma}^T\mathbf{n} \le 1$ and an ϵ-JF($\mathbf{n}$) trajectory $\mathbf{q}^\epsilon(\cdot)$ with the following properties:*

a. $q^\epsilon_m(1) > 0$.
b. *The times at which $\mathbf{q}^\epsilon(\cdot)$ is discontinuous (the jump times) all belong to the open interval* $(0, 1)$.
c. *At each jump time, exactly one of the components of $\mathbf{q}^\epsilon(\cdot)$ is discontinuous.*
d. *The arrival rate associated to $\mathbf{q}^\epsilon(\cdot)$ satisfies $\inf_t \lambda_j(t) > 0$ for all j.*

For the rest of the proof, we fix an ϵ-JF($\mathbf{n}$) trajectory with the properties in Lemma 7, together with the associated vector $\mathbf{n}$ and rate function $\boldsymbol{\lambda}(\cdot)$. We define $n = n_1 + \cdots + n_\ell$, which is the total number of jumps.[5]

We define $\Theta_0 = 0$, $\Theta_{n+1} = 1$, and for $k = 1, \ldots, n$, let Θ_k be the kth jump time. In particular,

$$0 = \Theta_0 < \Theta_1 < \cdots < \Theta_n < \Theta_{n+1} = 1. \tag{29}$$

We use j_k to refer to the queue at which the kth jump occurs and a_k to refer to the size of the kth jump. In particular, at time Θ_k, for $k = 1, \ldots, n$, $q_j(\cdot)$ is continuous for every $j \ne j_k$, and

$$q^\epsilon_{j_k}(\Theta_k) = q^\epsilon_{j_k}(\Theta_k^-) + a_k.$$

8.3. Defining Certain Constants

As in (16), we define

$$\overline{\mu} = 1 + \max_{\boldsymbol{\mu} \in \mathcal{M}} \|\boldsymbol{\mu}\| + \|\boldsymbol{\lambda}^*\| + \epsilon. \tag{30}$$

Moreover, similar to (13), we let

$$\gamma = \min_{j=1,\ldots,\ell} \gamma_j > 0. \tag{31}$$

By arguing as in Claim 1, we can and do assume, without loss of generality, that $\gamma \le 1$. We then define a positive constant

$$c = q^\epsilon_m(1), \tag{32}$$

and also let

$$d = \frac{1}{2}\min\left\{\frac{\gamma c}{4(1+4\overline{\mu})}, \quad \min_{k=0,\ldots,n}\{\Theta_{k+1}-\Theta_k\}, \quad \frac{\min_{k=1,\ldots,n} a_k}{1+2\overline{\mu}}\right\}. \tag{33}$$

(In case $n = 0$, the last term inside the brackets is taken to be zero.)

8.4. Defining the Stochastic Arrivals

Let us fix some constants $t_0 \geq 0$ and $T > 0$ and keep them fixed until the end of the proof in Section 8.7. In this section, we define a stochastic arrival process over an interval of the form $[t_0, t_0 + T)$, thus constructing what we call an *episode* of the overall process. Later, in Section 8.7, we concatenate multiple episodes to construct the stochastic process over the entire timeline $[0, \infty)$.

Consider the arrival rate function $\boldsymbol{\lambda}(\cdot)$ of the ϵ-JF($\mathbf{n}$) trajectory; in particular, $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}^*\| \leq \epsilon$. The arrival rate vector for the stochastic process during the episode is a time-scaled (by a factor of T) and shifted (by t_0) version of $\boldsymbol{\lambda}(\cdot)$. More concretely, we let

$$\overline{\boldsymbol{\lambda}}(t) = \boldsymbol{\lambda}\left(\frac{t-t_0}{T}\right), \qquad t \in [t_0, t_0+T). \tag{34}$$

Clearly,

$$\|\overline{\boldsymbol{\lambda}}(t) - \boldsymbol{\lambda}^*\| \leq \epsilon, \qquad \forall t \in [t_0, t_0+T). \tag{35}$$

Furthermore, it follows from Lemma 7(d) that $\inf_t \overline{\lambda}_j(t) > 0$ for every j.

We now digress to introduce certain constants that are used to specify the exact form of the distribution of the arrival processes. For any $\alpha > 0$, we let

$$\sigma(\alpha) = \int_{\overline{\mu}}^{\infty} x^{-(1+\alpha)} \log(x+1)\, dx, \tag{36}$$

where $\overline{\mu}$ is defined in (30). Then, for any $\alpha > 0$, we have

$$\frac{\sigma(\alpha)}{\sigma(1+\alpha)} = \frac{\int_{\overline{\mu}}^{\infty} x \cdot x^{-(2+\alpha)} \log(x+1)\, dx}{\int_{\overline{\mu}}^{\infty} x^{-(2+\alpha)} \log(x+1)\, dx} \geq \frac{\overline{\mu}\int_{\overline{\mu}}^{\infty} x^{-(2+\alpha)} \log(x+1)\, dx}{\int_{\overline{\mu}}^{\infty} x^{-(2+\alpha)} \log(x+1)\, dx} = \overline{\mu}. \tag{37}$$

Definition 8 (Arrivals During an Episode). Given $T > 0$ and $t_0 \geq 0$, we define the arrival process over the interval $t \in [t_0, t_0 + T)$ (which we call an episode) as follows:

a. If $\gamma_j = \infty$, then $A_j(t) = \overline{\lambda}_j(t)$ (deterministically).

b. If $\gamma_j < \infty$, then $A_j(t)$ is a random variable with probability density function

$$f_{A_j(t)}(x) = \frac{\overline{\lambda}_j(t)}{\sigma(\gamma_j)} \cdot x^{-(2+\gamma_j)} \log(x+1)\mathbb{1}(x \geq \overline{\mu}) + \left(1 - \frac{\overline{\lambda}_j(t)\sigma(1+\gamma_j)}{\sigma(\gamma_j)}\right)\delta(x), \tag{38}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, $\delta(\cdot)$ is Dirac's delta function, and $\sigma(\cdot)$ is as defined in (36).

c. The $A_j(t)$ for different j and t are independent.

We refer to $\{\mathbf{A}(t) : t \in [t_0, t_0+T)\}$ as an *episode-adjusted* arrival process.

From (37), we obtain $\sigma(\gamma_j)/\sigma(1+\gamma_j) \geq \overline{\mu} \geq \overline{\lambda}_j(t)$, where the last inequality is due to (30) and (35). Therefore, the coefficient of the delta function is nonnegative, and we have a well-defined distribution. It is also easy to verify that $\int_0^\infty f_{A_j(t)}(x)\, dx = 1$. Furthermore, from (35), $\overline{\lambda}_j(t)$ is bounded above. It follows that each $A_j(t)$ is dominated by a random variable $\overline{A}_j$ with tail exponent γ_j, and consequently, the process also has tail exponent γ_j. Moreover, $\mathbb{E}[A_j(t)] = \int_0^\infty x f_{A_j(t)}(x)\, dx = \overline{\lambda}_j(t)$. Therefore, using again (35), we have $\|\mathbb{E}[\mathbf{A}(t)] - \boldsymbol{\lambda}^*\| \leq \epsilon$, for all $t \in [t_0, t_0+T)$. In particular, the process $\mathbf{A}(\cdot)$ belongs to the class $\mathcal{A}_\epsilon(\boldsymbol{\gamma}; \boldsymbol{\lambda}^*)$ as desired.

8.5. Probabilistic Analysis

We aim to show that, during an episode and with significant probability, the queue vector process $\mathbf{Q}(\cdot)$ stays close to a scaled version of the ϵ-JF trajectory, that is, that

$$\mathbf{Q}(t) \approx T\mathbf{q}^\epsilon\left(\frac{t-t_0}{T}\right), \qquad \forall t \in [t_0, t_0+T).$$

To accomplish this, we consider each interval of the form $[t_0+\Theta_k T, t_0+\Theta_{k+1}T)$ for $k=0,1,\ldots,n$ and show that there is a substantial probability that the arrival process we have defined has the following properties: (a) as long as $k>0$, it has a jump in a small segment in the beginning of the interval, and (b) it has small fluctuations in the rest of the interval. We define events that capture these two properties.

For $k=1,\ldots,n$, let $\mathbf{B}_k$ be the cumulative arrival vector over the interval $[t_0+\Theta_k T, t_0+\Theta_k T+dT)$, that is,[6]

$$\mathbf{B}_k = \sum_{t=t_0+\Theta_k T}^{t_0+\Theta_k T+dT-1} \mathbf{A}(t). \tag{39}$$

We let $\mathcal{E}_k^{\text{jump}}$ be the event that $\mathbf{B}_k$ emulates the jump in $\mathbf{q}^\epsilon(\cdot)$ at time Θ_k scaled by T, that is,

$$\mathcal{E}_k^{\text{jump}} = \{\|\mathbf{B}_k - Ta_k \boldsymbol{e}_{j_k}\| \le dT(1+2\overline{\mu})\}, \tag{40}$$

where j_k is defined as the index of the queue at which the kth jump takes place and d is the constant defined in (33).

Lemma 8. *There exist $\psi\in(0,1)$ and $\overline{T}_1>0$ such that, if $T\ge\overline{T}_1$, then for $k=1,\ldots,n$ and for the episode-adjusted arrival process over an episode $[t_0,t_0+T)$, we have $\mathbb{P}(\mathcal{E}_k^{\text{jump}})\ge\psi T^{-\gamma_{j_k}}\log T$.*

The proof is given in Appendix C.2.

Recall now that the function $\boldsymbol{\lambda}(\cdot)$ driving the trajectory $\mathbf{q}^\epsilon(\cdot)$ is piecewise constant with a finite number of pieces. We use r to denote the number of such pieces. We then proceed to define certain small fluctuation events. For $k=0,1,\ldots,n$, we let $\mathcal{E}_k^{\text{fluc}}$ be the event that the cumulative fluctuations of $\mathbf{A}(\cdot)$ over the interval $[t_0+\Theta_k T+dT, t_0+\Theta_{k+1}T)$ are small, that is,

$$\mathcal{E}_k^{\text{fluc}} = \left\{ \left\| \sum_{\tau=t_0+\Theta_k T+dT}^{t} (\mathbf{A}(\tau)-\overline{\boldsymbol{\lambda}}(\tau)) \right\| \le \frac{\gamma c T}{32Cr}, \quad \forall t\in[t_0+\Theta_k T+dT, t_0+\Theta_{k+1}T) \right\}. \tag{41}$$

Lemma 9. *There exists a $\overline{T}_2\ge 0$ such that, if $T\ge\overline{T}_2$, then for $k=0,1,\ldots,n$ and for the episode-adjusted process over an episode $[t_0,t_0+T)$, we have $\mathbb{P}(\mathcal{E}_k^{\text{fluc}})\ge 1/2$.*

The proof is given in Appendix C.3.

Note that each one of the events $\mathcal{E}_k^{\text{jump}}$ for $k=1,\ldots,n$ and $\mathcal{E}_k^{\text{fluc}}$ for $k=0,\ldots,n$ is determined by the arrival process during a particular interval and all of these intervals are disjoint. Thus, the independence assumption on the arrival process implies that all of these events are independent. Therefore, if $T\ge\max\{\overline{T}_1,\overline{T}_2\}$, then

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1^{\text{jump}},\ldots,\mathcal{E}_n^{\text{jump}},\mathcal{E}_0^{\text{fluc}},\ldots,\mathcal{E}_n^{\text{fluc}}) &= \prod_{k=1}^n \mathbb{P}(\mathcal{E}_k^{\text{jump}})\cdot\prod_{k=0}^n \mathbb{P}(\mathcal{E}_k^{\text{fluc}}) \\
&\ge \prod_{k=1}^n \psi T^{-\gamma_{j_k}}\log T\cdot\prod_{k=0}^n \frac{1}{2} \\
&= \frac{1}{2}\left(\frac{\psi\log T}{2}\right)^n T^{-\sum_{k=1}^n \gamma_{j_k}} \\
&= \frac{1}{2}\left(\frac{\psi\log T}{2}\right)^n T^{-\boldsymbol{\gamma}^T\mathbf{n}} \\
&\ge \frac{1}{2}\left(\frac{\psi}{2}\right)^n T^{-\boldsymbol{\gamma}^T\mathbf{n}}\log T \\
&\ge \frac{1}{2}\left(\frac{\psi}{2}\right)^{1/\gamma} T^{-\boldsymbol{\gamma}^T\mathbf{n}}\log T \\
&\ge \frac{1}{2}\left(\frac{\psi}{2}\right)^{1/\gamma} T^{-1}\log T,
\end{aligned} \tag{42}$$

where the first inequality follows from Lemmas 8 and 9. The second inequality is because $n\ge 1$ and (without loss of generality) $\log T\ge 1$. The third inequality is due to $n\gamma=\sum_{j=1}^\ell n_j\gamma\le\sum_{j=1}^\ell n_j\gamma_j\le 1$ so that $n\le 1/\gamma$ together with the fact $\psi\le 1$ (see Lemma 8). The last inequality is again because $\boldsymbol{\gamma}^T\mathbf{n}\le 1$.

8.6. $\mathbb{E}[Q_m]$ is Large at the End of an Episode

We are now ready to argue that, if the events $\mathcal{E}_1^{\text{jump}},\dots,\mathcal{E}_n^{\text{jump}}$ and $\mathcal{E}_0^{\text{fluc}},\dots,\mathcal{E}_n^{\text{fluc}}$ occur, then, over an episode $[t_0,t_0+T)$, the process $\mathbf{Q}(\cdot)$ stays close to the suitably scaled ϵ-JF trajectory. As a consequence, the value of Q_m at the end of the episode becomes of order cT, where $c=q_m^{\epsilon}(1)>0$ as in (32). This argument is entirely deterministic. It is carried out in the course of the proof of the next lemma (in Appendix C.4) and relies on the sensitivity bound in Theorem 4.

Lemma 10. *There exists a $\overline{T}_3\geq 0$ such that, if $T\geq\overline{T}_3$, then the following holds. If a sample path of the episode-adjusted process over the episode $[t_0,t_0+T)$ satisfies the events $\mathcal{E}_1^{\text{jump}},\dots,\mathcal{E}_n^{\text{jump}}$ and $\mathcal{E}_0^{\text{fluc}},\dots,\mathcal{E}_n^{\text{fluc}}$ and if $\|\mathbf{Q}(t_0)\|\leq cT/5$, then $Q_m(t_0+T)\geq cT/2$.*

Let $\rho=(\psi/2)^{1/\gamma}/4$ and $\overline{T}=\max\{2,\overline{T}_1,\overline{T}_2,\overline{T}_3\}$, where $\overline{T}_1$, $\overline{T}_2$, and $\overline{T}_3$ are the constants in Lemmas 8–10, respectively. We note that all of these constants, $\overline{T}_1$, $\overline{T}_2$, $\overline{T}_3$, and therefore $\overline{T}$ as well, are defined in terms of general parameters and properties of the particular ϵ-JF trajectory and are deterministic. Lemma 10 and (42) imply that, for an episode (t_0,t_0+T) with $T\geq\overline{T}$ and initialized so that $\|\mathbf{Q}(t_0)\|\leq cT/5$, we have

$$\mathbb{P}\left(Q_m(t_0+T)\geq\frac{cT}{2}\right)\geq\mathbb{P}\left(\mathcal{E}_1^{\text{jump}},\dots,\mathcal{E}_n^{\text{jump}},\mathcal{E}_0^{\text{fluc}},\dots,\mathcal{E}_n^{\text{fluc}}\right)\geq 2\rho\,T^{-1}\log T, \tag{43}$$

which implies that

$$\mathbb{E}[Q_m(t_0+T)]\geq\frac{cT}{2}\,\mathbb{P}\left(Q_m(t_0+T)\geq\frac{cT}{2}\right)\geq\frac{cT}{2}\cdot 2\rho T^{-1}\log T=\rho c\log T. \tag{44}$$

8.7. Concatenating Episodes over the Entire Timeline

So far, we have defined and studied an arrival process over an episode $[t_0,t_0+T)$. We now concatenate a sequence of such episodes of increasing duration, which defines an arrival process over an infinite timeline.

We define times $T_0,T_1,T_2,\dots,$ and arrival processes for the intervals $[T_i,T_{i+1})$, recursively, as follows. We let $T_0=0$ and $T_1=\overline{T}$, where $\overline{T}$ is defined in the last paragraph of Section 8.6. We also let $\mathbf{A}(\cdot)$ for $t\in[T_0,T_1)$ be the corresponding episode-adjusted process as in Definition 8. Suppose now that we have defined T_i for some $i\geq 1$ as well as the arrival process for $t\in[0,T_i)$. We then let

$$T_{i+1}=T_i+\max\left\{T_i,\frac{10\,\mathbb{E}[\|\mathbf{Q}(T_i)\|]}{c}\right\}. \tag{45}$$

Finally, we define the arrival process over the episode $[T_i,T_{i+1})$ to be the corresponding episode-adjusted process. With this recursion, the arrival process is now well-defined for all times $t\geq 0$.

Note that $\mathbb{E}[\|\mathbf{Q}(T_i)\|]\leq\mathbb{E}\left[\sum_{t=0}^{T_i-1}\|\mathbf{A}(t)\|\right]=\sum_{t=0}^{T_i-1}\mathbb{E}[\|\mathbf{A}(t)\|]\leq T_i\ell(\|\boldsymbol{\lambda}^*\|+\epsilon)<\infty$. This guarantees that all T_i are finite and we have an infinite number of episodes. Moreover, note that, for $i=1,2,\dots,$ we have $T_{i+1}\geq 2T_i$. As a result, $T_i\geq 2^{i-1}\overline{T}$, and also,

$$\log(T_{i+1}-T_i)\geq\log T_i\geq i\log 2, \tag{46}$$

where the last inequality is because $T_i\geq 2^{i-1}\overline{T}\geq 2^i$.

We define the event

$$\mathcal{B}_i=\left\{\|\mathbf{Q}(T_i)\|\leq\frac{(T_{i+1}-T_i)c}{5}\right\},$$

and use the Markov inequality to obtain

$$\mathbb{P}(\mathcal{B}_i)\geq 1-\frac{\mathbb{E}[\|\mathbf{Q}(T_i)\|]}{(T_{i+1}-T_i)c/5}\geq 1-\frac{\mathbb{E}[\|\mathbf{Q}(T_i)\|]}{(10\mathbb{E}[\|\mathbf{Q}(T_i)\|])/5}=\frac{1}{2}, \tag{47}$$

where the second inequality is due to (45).

Recall now that Inequality (44), which is about an episode of length T, makes use of the assumption $\|\mathbf{Q}(t_0)\|\leq cT/5$, where t_0 is the start time of the episode. According to (47), this assumption is satisfied at the start time of the episode $[T_i,T_{i-1})$ with probability at least $1/2$. By interpreting (44) as a statement about conditional expectations and with t_0 and t_0+T replaced by T_i and T_{i+1}, respectively, we obtain

$$\mathbb{E}[Q_m(T_{i+1})]\geq\mathbb{P}(\mathcal{B}_i)\cdot\mathbb{E}[Q_m(T_{i+1})\,|\,\mathcal{B}_i]\geq\frac{1}{2}\cdot\rho c\,\log\,(T_{i+1}-T_i)\geq\frac{\rho c\,i\log 2}{2},$$

where the last inequality follows from (46). Therefore, $\mathbb{E}[Q_m(T_i)]$ grows unbounded as i increases. Consequently, under the arrival process that we constructed, queue m is not delay stable. This conclusion is obtained for any positive choice of ϵ, no matter how small, and establishes that queue m is not RDS. This completes the proof of the second direction of Theorem 1.

9. Proof of Theorem 2

9.1. Proof of the First Direction

Let us fix some $\epsilon > 0$. To establish one direction of the result, we assume that the ϵ-JF($\boldsymbol{\gamma}$) condition holds for every $\boldsymbol{\gamma} \in \Gamma$. We show that there exists a special ϵ-Lyapunov function.

Let $\mathcal{N}$ be the set of all nonnegative integer vectors $\mathbf{n}$ such that $n_j = 0$ for $j > h$; that is, we allow arbitrarily many jumps at the heavy-tailed queues and no jumps at the light-tailed ones. As in Definition 6, for any nonnegative integer vector $\mathbf{n}$, let $W(\mathbf{n})$ be the set of all points in $\mathbb{R}_+^\ell$ that are reachable by ϵ-JF($\mathbf{n}$) trajectories. Let $W = \cup_{\mathbf{n}\in\mathcal{N}} W(\mathbf{n})$ and consider a Lyapunov function $V(\cdot)$ equal to the distance from W, that is, $V(\mathbf{x}) = d(\mathbf{x}, W)$, for any $\mathbf{x} \in \mathbb{R}_+^\ell$. We show that this Lyapunov function has the desired properties.

The distance function is clearly Lipschitz continuous with a Lipschitz constant equal to one, which implies the first property in the definition of special ϵ-Lyapunov functions.

For the second property, Lemma 4(b) applies and shows that each set $W(\mathbf{n})$ is ϵ-invariant. It can be seen that the union, W, of the ϵ-invariant sets $W(\mathbf{n})$ is also ϵ-invariant. It then follows from Lemma 4(a) that W is ϵ-attracting. This proves the second property in Definition 5.

Note that every $\mathbf{n} \in \mathcal{N}$ satisfies the inequality $\boldsymbol{\gamma}^T\mathbf{n} \le 1$ for some $\boldsymbol{\gamma} \in \Gamma$. Because the ϵ-JF($\boldsymbol{\gamma}$) condition holds for every $\boldsymbol{\gamma} \in \Gamma$, it follows that every ϵ-JF($\mathbf{n}$) trajectory with $\mathbf{n} \in \mathcal{N}$ satisfies $q_m(t) = 0$ for all t. Hence, $q_m = 0$ for all $\mathbf{q} \in W$. Furthermore, because $\mathbf{0} \in W$, we have $V(\mathbf{0}) = 0$. This establishes the third property in Definition 5.

Finally, W is closed under jumps along coordinates associated with heavy-tailed arrivals. Therefore, $V(\cdot)$ is nonincreasing along those directions, and the fourth property in Definition 5 follows. Thus, $V(\cdot)$ has all the required properties of special ϵ-Lyapunov functions. This completes the proof of one direction of the theorem.

9.2. Proof of the Reverse Direction

We continue with the proof of the reverse direction. We fix some $\epsilon > 0$ and assume that there exists a special ϵ-Lyapunov function $V(\cdot)$ and let $W = \{\mathbf{x} \in \mathbb{R}_+^\ell \,|\, V(\mathbf{x}) = 0\}$. The argument rests on the ϵ-invariance of W, which, in turn, relies on some properties of the MW dynamics that we discuss next.

For any $\boldsymbol{\lambda}, \mathbf{x} \in \mathbb{R}_+^\ell$, let

$$\boldsymbol{\xi}_{\boldsymbol{\lambda}}(\mathbf{x}) = \dot{\mathbf{q}}(0), \tag{48}$$

where $\mathbf{q}(\cdot)$ is the fluid trajectory corresponding to arrival rate $\boldsymbol{\lambda}$ and initialized with $\mathbf{q}(0) = \mathbf{x}$. In view of (9), we have $\boldsymbol{\xi}_{\boldsymbol{\lambda}}(\mathbf{x}) \in \overline{\mathcal{D}}_{\boldsymbol{\lambda}}(\mathbf{x})$. Moreover, it is shown in lemma 2(a) of Sharifnassab et al. (2019) that $\boldsymbol{\xi}_{\boldsymbol{\lambda}}(\mathbf{x})$ has the minimum norm among all vectors in $\overline{\mathcal{D}}_{\boldsymbol{\lambda}}(\mathbf{x})$, that is,

$$\boldsymbol{\xi}_{\boldsymbol{\lambda}}(\mathbf{x}) = \operatorname*{argmin}_{\boldsymbol{\nu}\in\overline{\mathcal{D}}_{\boldsymbol{\lambda}}(\mathbf{x})} \|\boldsymbol{\nu}\|, \qquad \forall \mathbf{x} \in \mathbb{R}_+^\ell. \tag{49}$$

Here, the minimizer is unique.

Given a closed and convex set $\mathcal{A} \subset \mathbb{R}^\ell$ and a point $\mathbf{x} \in \mathbb{R}^\ell$, we denote by $\pi_{\mathcal{A}}(\mathbf{x})$ the projection of $\mathbf{x}$ on $\mathcal{A}$, defined as the point in $\mathcal{A}$ that is closest to $\mathbf{x}$. With this terminology, $\boldsymbol{\xi}_{\boldsymbol{\lambda}}(\mathbf{x})$ is the projection $\pi_{\overline{\mathcal{D}}_{\boldsymbol{\lambda}}(\mathbf{x})}(\mathbf{0})$ of the zero vector on the set $\overline{\mathcal{D}}_{\boldsymbol{\lambda}}(\mathbf{x})$.

In what follows, we also make use of an elementary property of projections: if $\mathcal{A}$ is a closed convex set, $\mathbf{b}$ is some vector, and $\mathcal{B} = \mathcal{A} + \mathbf{b}$, then

$$\|\pi_{\mathcal{A}}(\mathbf{x}) - \pi_{\mathcal{B}}(\mathbf{x})\| \le \|\mathbf{b}\|. \tag{50}$$

As a consequence of this, for any $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, and $\mathbf{x}$ in $\mathbb{R}_+^\ell$,

$$\begin{aligned}\|\boldsymbol{\xi}_{\boldsymbol{\lambda}_1}(\mathbf{x}) - \boldsymbol{\xi}_{\boldsymbol{\lambda}_2}(\mathbf{x})\| &= \|\pi_{\overline{\mathcal{D}}_{\boldsymbol{\lambda}_1}(\mathbf{x})}(\mathbf{0}) - \pi_{\overline{\mathcal{D}}_{\boldsymbol{\lambda}_2}(\mathbf{x})}(\mathbf{0})\| \\ &= \|\pi_{\overline{\mathcal{D}}_{\boldsymbol{\lambda}_1}(\mathbf{x})}(\mathbf{0}) - \pi_{\overline{\mathcal{D}}_{\boldsymbol{\lambda}_1}(\mathbf{x})+\boldsymbol{\lambda}_2-\boldsymbol{\lambda}_1}(\mathbf{0})\| \\ &\le \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|,\end{aligned} \tag{51}$$

where the second equality is because

$$\overline{\mathcal{D}}_{\boldsymbol{\lambda}_2}(\mathbf{x}) = \boldsymbol{\lambda}_2 - \overline{\mathcal{M}}(\mathbf{x}) = \boldsymbol{\lambda}_1 - \overline{\mathcal{M}}(\mathbf{x}) + \boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1 = \overline{\mathcal{D}}_{\boldsymbol{\lambda}_1}(\mathbf{x}) + \boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1,$$

and the inequality follows from (50).

Lemma 11. *The set W is ϵ-invariant.*

Proof. Because V is a special ϵ-Lyapunov function, it is Lipschitz continuous with Lipschitz constant one. Let $\boldsymbol{\lambda} \in \mathbb{R}_+^\ell$ be such that $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\| \le \epsilon$ and consider fluid trajectories $\mathbf{q}(\cdot)$ and $\mathbf{p}(\cdot)$ corresponding to arrival rates $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^*$, respectively, initialized with the same nonnegative vector $\mathbf{q}(0) = \mathbf{p}(0) \notin W$. Then,

$$\begin{aligned}
\dot{V}(\mathbf{q}(t))\Big|_{t=0} &= \lim_{\delta \downarrow 0} \frac{V(\mathbf{q}(\delta)) - V(\mathbf{q}(0))}{\delta} \\
&= \lim_{\delta \downarrow 0} \frac{V(\mathbf{p}(\delta)) + V(\mathbf{q}(\delta)) - V(\mathbf{p}(\delta)) - V(\mathbf{q}(0))}{\delta} \\
&\le \lim_{\delta \downarrow 0} \frac{[V(\mathbf{p}(\delta)) + \|\mathbf{q}(\delta) - \mathbf{p}(\delta)\|] - V(\mathbf{q}(0))}{\delta} \\
&= \lim_{\delta \downarrow 0} \left(\frac{V(\mathbf{p}(\delta)) - V(\mathbf{p}(0))}{\delta} + \frac{\|\mathbf{q}(\delta) - \mathbf{p}(\delta)\|}{\delta} \right) \\
&\le \lim_{\delta \downarrow 0} \left(\frac{V(\mathbf{p}(\delta)) - V(\mathbf{p}(0))}{\delta} + \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\| \right) \\
&= \dot{V}(\mathbf{p}(t))|_{t=0} + \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\| \\
&\le -\epsilon + \epsilon \\
&= 0,
\end{aligned} \tag{52}$$

where the first inequality is because V has a Lipschitz constant equal to one, the third equality is due to $\mathbf{q}(0) = \mathbf{p}(0)$, and the second inequality follows from (51). The last inequality follows from the second property of special ϵ-Lyapunov functions in Definition 5 and the assumption $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\| \le \epsilon$.

This argument shows that, for any $\boldsymbol{\lambda} \in \mathbb{R}_+^\ell$ with $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\| \le \epsilon$ and for any fluid trajectory $\mathbf{q}(\cdot)$ corresponding to arrival rate $\boldsymbol{\lambda}$, the distance from the set W cannot increase. In particular, if $\mathbf{q}(\cdot)$ is initialized with $\mathbf{q}(0) \in W$, it must stay in W. Therefore, using the terminology in Definition 7, W is ϵ-invariant. □

From the third property in the definition of special ϵ-Lyapunov functions, we have $V(\mathbf{0}) = 0$ and, therefore, $\mathbf{0} \in W$. Thus, every ϵ-JF($\mathbf{n}$) trajectory starts in W. From the fourth property in the definition of special ϵ-Lyapunov functions, W is closed with respect to positive jumps along the coordinates associated with heavy-tailed arrivals. Using also Lemma 11 for the times between jumps, we see that, for any $\mathbf{n} \in \mathcal{N}$, every ϵ-JF($\mathbf{n}$) trajectory stays in W. Equivalently, for any $\boldsymbol{\gamma} \in \Gamma$, every ϵ-JF($\boldsymbol{\gamma}$) trajectory stays in W. Finally, employing again the third property of special ϵ-Lyapunov functions, we have $q_m = 0$ for all $\mathbf{q} \in W$. This implies that $q_m(t) = 0$ for all ϵ-JF($\boldsymbol{\gamma}$) trajectories $\mathbf{q}(\cdot)$ with $\boldsymbol{\gamma} \in \Gamma$ and all $t \ge 0$. This establishes the ϵ-JF($\boldsymbol{\gamma}$) condition for all $\boldsymbol{\gamma} \in \Gamma$ and completes the proof of Theorem 2.

9.3. Proof of Corollary 1

For part (a), fix some $\epsilon > 0$ and suppose that there exists a special ϵ-Lyapunov function. Theorem 2 implies that the ϵ-JF($\boldsymbol{\gamma}$) condition holds for every $\boldsymbol{\gamma} \in \Gamma$. It then follows from Theorem 1 that queue m is RDS for every $\boldsymbol{\gamma} \in \Gamma$.

For part (b), we fix some $\epsilon > 0$ and assume that queue m is ϵ-RDS for all $\boldsymbol{\gamma} \in \Gamma$. In particular, queue m is RDS, and Theorem 1 implies that the ϵ'-JF($\boldsymbol{\gamma}$) condition holds for some $\epsilon' > 0$. However, a close inspection of the proof of the reverse part of Theorem 1 reveals that we can in fact choose ϵ' to be the same as ϵ. Thus, the ϵ-JF($\boldsymbol{\gamma}$) condition holds, and Theorem 2 implies that there exists a special ϵ-Lyapunov function.

10. Discussion

In this section, we summarize some key points and conclude with a few open questions.

10.1. Framing and Results

We have addressed the problem of delay stability for a class of queueing networks that operate under the max-weight scheduling policy when some arrival processes are heavy-tailed and some are light-tailed. The overall purpose was to develop conditions for delay stability in terms of fluid-like models. However, as illustrated by

the example in Section 3, delay instability can be the result of multiple coordinated large jumps. The probabilities of such large jumps are, in turn, affected by the tail exponents of the arrival processes. Given that traditional fluid models are oblivious to the tail exponents, we introduce JF models, a generalization that allows for jumps along the coordinates associated with heavy-tailed flows subject to a budget on the number of jumps with the budget being determined by the tail exponents.

At the same time, it became clear that tight conditions for delay stability that do not depend on the details of the arrival distributions are only possible under a suitable robust formulation with respect to both the arrival rates and the arrival process distributions. With a careful choice of definitions, we were finally able to establish necessary and sufficient conditions for robust delay stability in terms of ϵ-JF models.

In Section 3, we also discuss a related, so-called ZF condition. The ZF condition essentially examines fluid trajectories that start at zero and involve a single jump and leads to a necessary condition for delay stability (Markakis et al. 2018), but the question whether it can also form the basis for a sufficient condition was open. Our results show that, in order to obtain necessary and sufficient conditions, we need to examine a richer set of trajectories that involve multiple jumps.

Finally, earlier works (Markakis 2013, Markakis et al. 2018) show that Lyapunov functions with certain structural properties can yield sufficient conditions for delay stability. But it was not clear if and when delay stability is equivalent to the existence of such Lyapunov functions. Our results make progress toward establishing the completeness of such a Lyapunov-based methodology for the regime in which the heavy-tailed flows can have arbitrarily small tail exponents.

Our RJF condition is difficult to test for general networks. In some sense, this reflects the intrinsic complexity of the (robust) delay stability problem. The Lyapunov-based condition also appears to be hard to test for general networks.

10.2. Alternative Formulations

Given the complexity of the RJF condition, it is natural to inquire about simpler alternatives. For example, is it possible to obtain tight delay stability results (without robustness) if we consider JF models with a constant rate $\boldsymbol{\lambda}(\cdot)$? In the same spirit, can we restrict to the case in which all jumps in ϵ-JF models take place at the same time? Might it be easier to consider concrete arrival processes instead of focusing on delay stability for all arrival processes with given exponents?

For all three of these questions, the answer is negative. We discuss such variations and related (counter)examples in Appendix A.

10.3. Open Problems

We collect here a few open problems and possible future research directions.

a. Can the results be generalized to other scheduling policies, for example, MW-α policies (Neely 2010), an extension of the MW policy considered in this paper, or more generally to other stochastic networks whose stability has been studied using fluid models? One obstacle here is that our main result relies heavily on a particular fluctuation bound, which is established specifically for the MW dynamics (Sharifnassab et al. 2020). However, progress may be possible if we rely on alternative stochastic bounding techniques.

b. Can we identify some special classes of networks for which our criteria (either the RJF condition or the Luyapunov-based condition) can be tested in polynomial time?

c. Is the RJF condition, which involves time-varying $\boldsymbol{\lambda}(t)$ with $\boldsymbol{\lambda}(t) \approx \boldsymbol{\lambda}^*$ equivalent to a similar condition in which we only consider ϵ-JF trajectories with time-invariant $\boldsymbol{\lambda}(t) = \boldsymbol{\lambda}$ with $\boldsymbol{\lambda} \approx \boldsymbol{\lambda}^*$? See Appendix A for further discussion and some conjectures.

Acknowledgments

We are grateful to Dr. Bert Zwart for providing us with pointers to relevant works in the literature and for informative technical discussions. Various stages of this work were completed while A. Sharifnassab was affiliated with the Massachusetts Institute of Technology, Sharif University of Technology, and the University of Alberta.

Appendix A. Exploring Alternative Formulations

Our formulation involves rate-robustness (robustness with respect to variations in the arrival rate) as well as distributional robustness (by considering the worst case over all distributions with given tail exponents). It would be preferable to develop conditions that characterize delay stability for specific systems (with fixed arrival rates and arrival distributions). However, this seems to be impossible for reasons that become clearer in this appendix. In particular, the distributional robustness aspect appears to be inevitable as long as we are aiming at conditions that are both necessary and

sufficient; see Appendix A.4. For this reason, most of this appendix is devoted to exploring variants of rate robustness. In the interest of brevity, we keep the discussion informal without rigorous proofs.

A.1. Variations of Our Definitions

In this section, we present a number of variations to our definitions of RDS and the RJF condition. In later sections, we elaborate on their relations. Throughout this appendix, we assume that the tail exponent vector $\boldsymbol{\gamma}$ is fixed with every γ_j in $(0,\infty]$. We also fix some $\boldsymbol{\lambda}^* > 0$. The various definitions that we offer differ only with respect to the choice of allowed functions $\boldsymbol{\lambda}(\cdot)$.

Let F be a class of discrete-time functions $\boldsymbol{\lambda}(\cdot)$. We say that queue m is F-RDS if it obeys Definition 1 except that the allowed arrival rates $\mathbb{E}[\mathbf{A}(t)]$ are also required to belong to F.

We consider the following choices for F, leading to three alternative definitions of robust delay stability, namely, G-RDS, C-RDS, and 0-RDS:

G (general): Here, we impose no additional restrictions on $\mathbb{E}[\mathbf{A}(t)]$. Thus, G-RDS is identical to the RDS condition that we study.

C (constant): Here, we require $\mathbb{E}[\mathbf{A}(t)]$ to be constant. Effectively, we are considering small but constant perturbations of $\boldsymbol{\lambda}^*$.

0 (zero): Here, we require $\mathbb{E}[\mathbf{A}(t)]$ to be equal to $\boldsymbol{\lambda}^*$ for all times t.

We continue similarly to define variants of the RJF condition. Let F be a class of continuous-time functions $\boldsymbol{\lambda}(\cdot)$. The F-RJF condition is defined exactly as in Definition 4 except that we only consider ϵ-JF trajectories for which the rate function $\boldsymbol{\lambda}(\cdot)$ is also required to belong to the class F. We consider four possible choices for F, leading to four variants of the RJF condition, namely, UC-RJF, PC-RJF, C-RJF, 0-RJF:

UC (uniformly continuous): Here, we remove the requirement in Definition 3 that $\boldsymbol{\lambda}(\cdot)$ be piecewise constant. Instead, we require $\boldsymbol{\lambda}(\cdot)$ to be (i) piecewise continuous with a finite number of discontinuities and (ii) uniformly continuous on any interval in which it is continuous.

PC (piecewise constant): Here, $\boldsymbol{\lambda}(\cdot)$ is exactly as in Definition 3, and in particular, piecewise constant. Thus, the PC-RJF condition coincides with the RJF condition we are studying.

C (constant): Here, we require $\boldsymbol{\lambda}(\cdot)$ to be constant. Effectively, we are considering small but constant perturbations of $\boldsymbol{\lambda}^*$.

0 (zero): Here, we require $\boldsymbol{\lambda}(\cdot)$ to be equal to $\boldsymbol{\lambda}^*$ for all times t.

A.2. Relations Between Alternative Definitions

In this section, we explore the relation between F-RDS and F-RJF conditions for different choices of F; see Figure A.1 for a visual summary.

It is clear that, when we restrict to a smaller class, the RDS or RJF conditions are easier to satisfy. Thus,

$$\text{G-RDS} \Rightarrow \text{C-RDS} \Rightarrow \text{0-RDS},$$

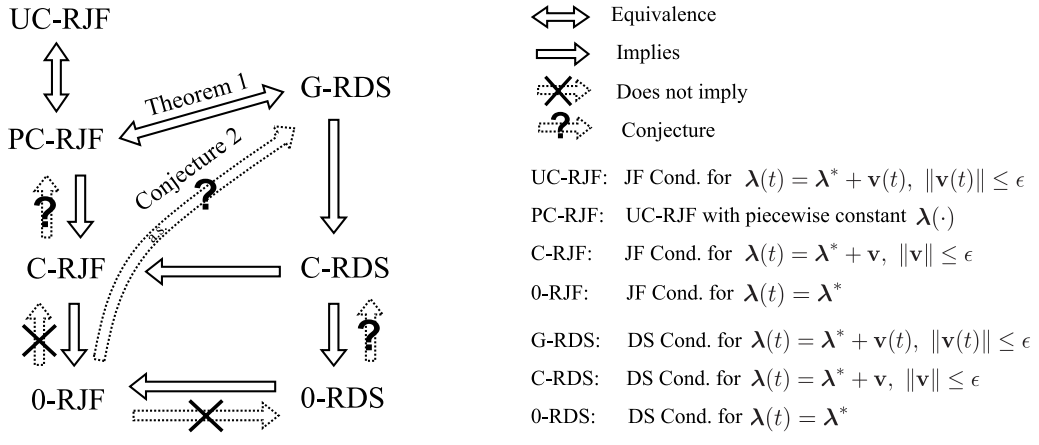
and

$$\text{UC-RJF} \Rightarrow \text{PC-RJF} \Rightarrow \text{C-RJF} \Rightarrow \text{0-RJF}.$$

Furthermore, Theorem 1 establishes that G-RDS is equivalent to PC-RJF.

A.2.1. PC-RJF ⇒ UC-RJF. An arbitrary (continuous-time) function $\boldsymbol{\lambda}(\cdot)$ in the class UC can be approximated by a piecewise constant function with finitely many pieces uniformly over a compact set. Furthermore, it can be shown that, if we

Figure A.1. Relation Between the Various Conditions



perturb by ϵ the vector $\boldsymbol{\lambda}(\cdot)$ that drives a JF-trajectory, the resulting trajectory is perturbed by at most ϵ over a time interval of length one. It follows that, if the UC-RJF condition fails, we can construct piecewise-constant approximations of $\boldsymbol{\lambda}(\cdot)$ that demonstrate that the PC-RJF condition also fails. Therefore, PC-RJF $\Rightarrow$ UC-RJF.

Taking Theorem 1 also into account, we see that all three conditions, G-RDS, PC-RJF, and UC-RJF, are equivalent. An alternative path to the same conclusion consists of modifying the proof in Section 8 and showing that G-RDS $\Rightarrow$ UC-RJF. This is possible, but quite tedious.

A.2.2. (C-RDS $\Rightarrow$ C-RJF) and (0-RDS $\Rightarrow$ 0-RJF). These two implications are true because the proof in Section 8 applies verbatim. Indeed, if we assume that C-RJF fails to hold, we start with a trajectory $\mathbf{q}(\cdot)$ that is driven by a constant rate $\boldsymbol{\lambda}$ and drives queue m to a positive value. The construction of the arrival process in Section 8.4 yields a process with a constant rate $\overline{\boldsymbol{\lambda}}$. Thus, the same proof establishes that failure of the C-RJF condition leads to failure of C-RDS; equivalently, C-RDS implies C-RJF. The argument that 0-RDS $\Rightarrow$ 0-RJF is the same.

A.2.3. 0-RJF $\not\Rightarrow$ 0-RDS. This fact exemplifies the difficulty of obtaining necessary and sufficient conditions in the absence of robustness considerations with respect to the arrival rates.

The argument is simple. Consider a single queue that is served at unit rate and let $\lambda^* = 1$. Suppose that the tail exponent is larger than one so that no jumps are allowed. In that case, there is only one possible JF trajectory, which obeys $\dot{q} = 1 - 1 = 0$. When initialized at zero, the JF trajectory stays at zero. Thus, the 0-RJF condition holds for the single queue of interest. On the other hand, as long as the arrivals are not deterministic, the stochastic system is marginally unstable, the expected queue length grows to infinity, and the 0-RDS condition does not hold.

This example involves a system operating at the boundary of its capacity region (marginally unstable). We can also construct simple examples (involving two queues) in which the system operates in the interior of the stability region, is stable and satisfies the 0-RJF condition but is not 0-RDS. Such an example (which we omit) involves a system that operates at the threshold between robust delay stability and instability.

A.2.4. 0-RJF $\not\Rightarrow$ C-RJF. This is again a simple observation. Consider the same single-queue system as in the previous paragraph with $\lambda^* = 1$. As long as the rate is fixed at one, the JF trajectory stays at zero, and the 0-RJF condition holds. On the other hand, a small constant perturbation that results in $\lambda > 1$ yields a divergent JF trajectory, and therefore, the C-RJF condition does not hold.

A.3. Conjectures and Open Problems

We list here a number of questions and conjectures.

A.3.1. 0-RDS $\stackrel{?}{\Rightarrow}$ C-RDS. We conjecture that, when $\boldsymbol{\lambda}^* > \mathbf{0}$,[7] 0-RDS implies C-RDS. Ultimately, this amounts to showing that the set of positive arrival rate vectors λ^* for which the system is delay stable (robustly over all distributions with given tail exponents) is open. The rationale behind this conjecture is that, in more standard settings (ordinary stability) the set of positive vectors $\boldsymbol{\lambda}^*$ that lead to a stable system is open.

A.3.2. C-RJF $\stackrel{?}{\Rightarrow}$ PC-RJF. We conjecture that this implication is true although we do not see how to establish it. If it is true, it would follow from the diagram in Figure A.1 that C-RJF and C-RDS are equivalent to G-RDS, UC-RJF, and PC-RJF.

An indirect approach to establishing the conjecture is to show that (i) C-RJF $\Rightarrow$ C-RDS and (ii) C-RDS $\Rightarrow$ G-RDS. However, this appears to be difficult. Our proof that PC-RJF implies G-RDS involves the set W of points reachable by ϵ-JF trajectories; see Definition 6. However, when we restrict $\boldsymbol{\lambda}(\cdot)$ to be constant, this set is no longer ϵ-invariant, and Lemma 4(ii) fails to go through.

A.3.3. Generic Considerations. A fundamental reason behind the mismatch between 0-RJF and G-RDS is that, at least for simple examples, the set of nonnegative nominal rates $\boldsymbol{\lambda}^*$ for which 0-RJF holds is closed, whereas the set of positive nominal rates $\boldsymbol{\lambda}^*$ for which G-RDS holds is open. It is conceivable, however, that one set is the closure of the other and the difference between the two sets is just a lower dimensional boundary. This leads us to the conjecture that 0-RJF and G-RDS are generically equivalent.

Hypothesis A.1. *Let us fix a network and some $\boldsymbol{\gamma}$. The set of nonnegative nominal arrival vectors $\boldsymbol{\lambda}^*$ for which the 0-RJF condition holds but G-RDS does not hold has zero Lebesgue measure.*

A.4. The Details of the Arrival Distribution May Matter

Our discussion so far has been about distributionally robust results, dealing with delay stability for all arrival distributions with the given tail exponents $\boldsymbol{\gamma}$. The reason for this was that JF models cannot take into account any further properties of these distributions.

Once we start inquiring about delay stability for a fixed, fully specified system, the situation is more complex: necessary and sufficient conditions for delay stability appear to be impossible. We illustrate the situation by stating a positive result and disciussing the obstacles in establishing a converse.

A.4.1. Delay Stability Implies the 0-RJF Condition Under a Regularity Assumption. Suppose that a particular system (with a constant arrival rate $\boldsymbol{\lambda}^*$ and given, i.i.d. arrival distributions) is delay stable. Suppose, furthermore, that the

distribution of each $A_j(t)$ satisfies (3) with γ replaced by the appropriate γ_j. Then, it can be shown that the 0-RJF condition holds. The argument involves similar ideas as the proof in Section 8. That is, we can show that the stochastic system can track an ϵ-JF trajectory with significant probability.

Note, however, that the 0-RJF condition does not imply delay stability,even under such a regularity assumption. The argument is the same as in our earlier example that showed that the 0-RJF condition does not imply the 0-RDS condition.

A.4.2. Without a Regularity Assumption, Delay Stability Need Not Imply the 0-RJF Condition. In contrast to the aforementioned result, we have strong reasons to conjecture that there exist systems that are delay stable, and yet, the 0-RJF condition fails to hold. The intuition behind this conjecture is as follows.

Consider a system with two heavy-tailed arrival streams together with some light-tailed ones. Suppose that the tail exponents of the heavy-tailed arrivals are larger than 1/3. We can arrange the system so that a JF trajectory drives the light-tailed queue of interest to a positive value if and only if we have one jump at each heavy-tailed queue, the two jump times are approximately equal, and the jump sizes are comparable (within a constant factor of each other). Such a system does not satisfy the 0-RJF condition.

As in the proof in Section 8, we might expect that the stochastic system can track this JF trajectory. However, we can arrange the arrival process distributions for the two heavy-tailed queues to be such that their supports are wide apart. For example, one distribution may be supported on integers of the form 10^{2i} and the other on integers of the form 10^{2i+1}. In that case, equal-size jumps are essentially impossible. As a consequence, the stochastic system should be unable to emulate the JF trajectory, the instability mechanism suggested by the JF trajectory need not be present, and the queue of interest may turn out to be delay stable.

A.5. The Timing of the Jumps

The definition of ϵ-JF trajectories allows for jumps at different times. On the other hand, our examples so far rely on jumps that happen simultaneously. This raises the question whether the RJF condition is equivalent to an analogous condition in which we only consider trajectories with simultaneous jumps. It turns out that this is not possible. We give an example with four queues in which an ϵ-JF trajectory drives a certain queue to a positive value, but this is only possible if we allow jumps to occur at different times.

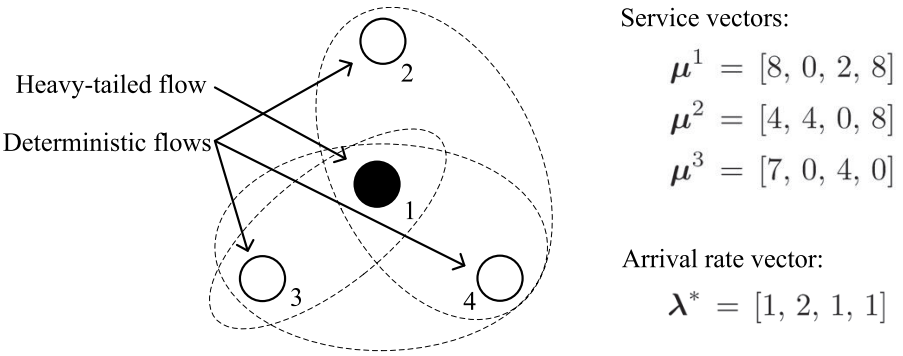
Consider the system in Figure A.2. The first queue receives heavy-tailed arrivals with $\gamma_1 = 1/2$, whereas the three other queues receive light-tailed arrivals. There are three possible service vectors as shown in the figure, and the arrival rate vector is $\boldsymbol{\lambda}^* = (1,2,1,1)$. Note that the condition $\boldsymbol{\gamma}^T\mathbf{n} \le 1$ allows up to two jumps at queue 1. If we restrict to simultaneous jumps, this essentially limits us to a single jump at queue 1.

Suppose that q_1 has a jump of size 27 at time 0. Then, the 0-JF trajectory is piecewise linear with breakpoints $\mathbf{q}(0) = (27,0,0,0)$, $\mathbf{q}(3) = (6,6,0,0)$, $\mathbf{q}(5) = (0,2,2,0)$, and $\mathbf{q}(9) = \mathbf{0}$. This can be easily verified by noticing that the MW policy chooses service vector $\boldsymbol{\mu}^1$ for $t \in [0,3)$ and service vector $\boldsymbol{\mu}^2$ for $t \in (3,5)$, and the service capacity is split between $\boldsymbol{\mu}^2$ and $\boldsymbol{\mu}^3$ with ratios 5/8 and 3/8 for $t \in (5,9)$. It then follows from the form of this piecewise linear fluid trajectory that, given a single jump at time 0, q_4 stays at zero for all subsequent times. We now argue that this is not the case if q_1 undergoes two jumps at different times.

Suppose that q_1 has a jump of size 27 at time 0 and a jump of size 2 at time 5. Let $\mathbf{p}(\cdot)$ be the associated jumping fluid trajectory. Then, right before time 5, we have $\mathbf{p}(5^-) = \mathbf{q}(5^-) = (0,2,2,0)$. Therefore, after the second jump, we have $\mathbf{p}(5) = (2,2,2,0)$. In this case, $\boldsymbol{\mu}^3$ is the dominant service vector for some positive time interval starting from time 5. Because q_4 receives no service under $\boldsymbol{\mu}^3$, it starts to build up and become positive.

In this example, the 0-RJF condition fails to hold, and the system is not 0-RDS. On the other hand, if we were to restrict to simultaneous jumps, we would not be able to tell that this is the case. Finally, using the same example, we see that we should also consider nonsimultaneous jumps when examining the C-RJF or PC-RJF conditions.

Figure A.2. A Network with Three Light-Tailed Queues and One Heavy-Tailed Queue, Which Demonstrates the Importance of the Timing of Multiple Jumps



Notes. In the example of this figure, if q_1 undergoes a single jump at time 0, q_4 stays at zero. However, two jumps in q_1 at suitably arranged times can result in a positive q_4.

Appendix B. Proofs of Lemmas for the First Direction of Theorem 1 (ϵ-JF $\Rightarrow$ RDS)

B.1. Proof of Theorem 4

We compare the fluid trajectory $\mathbf{q}(\cdot)$, which is initialized with $\mathbf{q}(0) \neq \mathbf{Q}(0)$, with another fluid trajectory $\tilde{\mathbf{q}}(\cdot)$, initialized with $\tilde{\mathbf{q}}(0) = \mathbf{Q}(0)$. From the triangle inequality,

$$\|\mathbf{Q}(t) - \mathbf{q}(t)\| \le \|\mathbf{Q}(t) - \tilde{\mathbf{q}}(t)\| + \|\tilde{\mathbf{q}}(t) - \mathbf{q}(t)\|.$$

We apply Theorem 3 to bound the first term on the right-hand side. Because of the nonexpansive property of the MW dynamics, we also have

$$\|\tilde{\mathbf{q}}(t) - \mathbf{q}(t)| \le \|\tilde{\mathbf{q}}(0) - \mathbf{q}(0)\| = \|\mathbf{Q}(0) - \mathbf{q}(0)\|,$$

and the result follows.

B.2. Proof of Lemma 1

Let us fix T throughout this proof. Recall the constant $\beta \in (1,2)$ introduced in the context of (19). For $j = 1, \ldots, \ell$, we define

$$\gamma_j' = \begin{cases} \gamma_j/\beta & \text{if } \gamma_j < \beta^3, \\ \beta^2 & \text{if } \gamma_j \ge \beta^3. \end{cases}$$

We then let $\boldsymbol{\gamma}' = (\gamma_1', \ldots, \gamma_\ell')$, and $\gamma' = \min_j \gamma_j'$. Note that, in all cases, we have $\gamma_j' \le \gamma_j/\beta$ so that $\gamma' \le \gamma/\beta$. As argued in Claim 1, we can and do (without loss of generality) assume that $\gamma = \min_j \gamma_j \le 1$, and thus, $\gamma' < 1$. Finally, in view of (19), it can be seen that, for any nonnegative integer vector $\mathbf{n}$,

$$\text{if} \quad \boldsymbol{\gamma}^T \mathbf{n} > 1, \quad \text{then} \quad (\boldsymbol{\gamma}')^T \mathbf{n} \ge \beta^2 > 1. \tag{B.1}$$

For every j and according to our definition (5) of the tail exponent of an arrival process, there is a random variable $\overline{A}_j$ that dominates $A_j(t)$ for all $t \ge 0$ and for which all moments of order less than $1 + \gamma_j$ are finite; see (3). We define

$$\Gamma_j = \mathbb{E}[\overline{A}_j^{1+\gamma_j'}], \qquad j = 1, \ldots, \ell, \tag{B.2}$$

which is finite because $\gamma_j' < \gamma_j$.

For $t = 0, \ldots, T-1$, and $j = 1, \ldots, \ell$, let

$$p_{j,t} = \mathbb{P}(A_j(t) > \theta_t).$$

For any j and $t \le T-1$, the Markov inequality yields

$$\begin{aligned} p_{j,t} &= \mathbb{P}(A_j(t) > \theta_t) \\ &\le \mathbb{P}(\overline{A}_j > \theta_t) \\ &= \mathbb{P}\left(\overline{A}_j^{1+\gamma_j'} > \theta_t^{1+\gamma_j'}\right) \\ &\le \frac{\mathbb{E}[\overline{A}_j^{1+\gamma_j'}]}{\theta_t^{1+\gamma_j'}} \\ &= \frac{\Gamma_j \eta^{1+\gamma_j'} \log^{1+\gamma_j'}(M+T-t)}{(M+T-t)^{1+\gamma_j'}}, \end{aligned} \tag{B.3}$$

where the last equality is due to the definitions of θ_t and Γ_j in (11) and (B.2), respectively.

Let $\phi = 1 - 2^{-1/\ell}$ and note that $0 < \phi < 1$. Because $0 < (1 - 1/\beta)\gamma' < \gamma_j'$ for every j, there exists some $M_1 \ge 1$ such that, if $M \ge M_1$, then

$$\Gamma_j \eta^{1+\gamma_j'} \log^{1+\gamma_j'} M \le \phi \cdot (\gamma_j' - (1 - 1/\beta)\gamma') \cdot M^{(1-1/\beta)\gamma'}, \quad j = 1, \ldots, \ell. \tag{B.4}$$

We fix such an M_1. Then, for $M \ge M_1$, we have

$$\begin{aligned} \sum_{t=0}^{T-1} p_{j,t} &\le \Gamma_j \eta^{1+\gamma_j'} \sum_{t=0}^{T-1} \frac{\log^{1+\gamma_j'}(M+T-t)}{(M+T-t)^{1+\gamma_j'}} \\ &\le \Gamma_j \eta^{1+\gamma_j'} \sum_{\tau=1}^{\infty} \frac{\log^{1+\gamma_j'}(M+\tau)}{(M+\tau)^{1+\gamma_j'}} \\ &\le \phi(\gamma_j' - (1-1/\beta)\gamma') \sum_{\tau=1}^{\infty} \frac{(M+\tau)^{(1-1/\beta)\gamma'}}{(M+\tau)^{1+\gamma_j'}} \\ &\le \phi(\gamma_j' - (1-1/\beta)\gamma') \int_M^{\infty} \frac{x^{(1-1/\beta)\gamma'}}{x^{1+\gamma_j'}} dx \\ &= \phi M^{(1-1/\beta)\gamma' - \gamma_j'}, \end{aligned} \tag{B.5}$$

where the first inequality is due to (B.3) and the third inequality follows from (B.4).

Let $X_{j,t}$ be the event $\{A_j(t) > \theta_t\}$. Recall that N_j stands for the number of jumps at queue j during the interval $[0,T)$. For any j and any nonnegative integer n, we have

$$\begin{aligned}\mathbb{P}(N_j = n) &\le \sum_{0\le\tau_1<\cdots<\tau_n\le T-1} \mathbb{P}(X_{j,\tau_1}\cap\cdots\cap X_{j,\tau_n})\\ &= \sum_{0\le\tau_1<\cdots<\tau_n\le T-1} p_{j,\tau_1}\cdots p_{j,\tau_n}\\ &\le \left(\sum_{t=0}^{T-1} p_{j,t}\right)^n\\ &\le \phi^n M^{(1-1/\beta)\gamma' n-\gamma'_j n}. \end{aligned} \tag{B.6}$$

Here, the first inequality follows from the union bound, the equality is from the independence of the events $X_{j,\tau_1},\ldots,X_{j,\tau_n}$, and the last inequality is due to (B.5). Therefore, for any $\mathbf{n}\in\mathbb{Z}_+^\ell$,

$$\begin{aligned}\mathbb{P}(\mathbf{N}=\mathbf{n}) &= \prod_{j=1}^{\ell}\mathbb{P}(N_j=n_j)\\ &\le \prod_{j=1}^{\ell}\phi^{n_j}M^{(1-1/\beta)\gamma' n_j}M^{-\gamma'_j n_j}\\ &= \phi^{|\mathbf{n}|}M^{(1-1/\beta)\gamma'|\mathbf{n}|-\mathbf{n}^T\boldsymbol{\gamma}'}, \end{aligned} \tag{B.7}$$

where $|\mathbf{n}| = n_1+\cdots+n_\ell$. If $\boldsymbol{\gamma}^T\mathbf{n}>1$, then (B.1) asserts that $\mathbf{n}^T\boldsymbol{\gamma}'\ge\beta^2$. As a result and because γ' is the smallest component of $\boldsymbol{\gamma}'$,

$$\left(1-\frac{1}{\beta}\right)\gamma'|\mathbf{n}| - \mathbf{n}^T\boldsymbol{\gamma}' \le \left(1-\frac{1}{\beta}\right)\mathbf{n}^T\boldsymbol{\gamma}' - \mathbf{n}^T\boldsymbol{\gamma}' = -\frac{\mathbf{n}^T\boldsymbol{\gamma}'}{\beta}\le -\beta. \tag{B.8}$$

Hence,

$$\begin{aligned}\mathbb{P}(\boldsymbol{\gamma}^T\mathbf{N}>1) &= \sum_{\mathbf{n}:\,\boldsymbol{\gamma}^T\mathbf{n}>1}\mathbb{P}(\mathbf{N}=\mathbf{n})\\ &\le \sum_{\mathbf{n}:\,\boldsymbol{\gamma}^T\mathbf{n}>1}\phi^{|\mathbf{n}|}M^{(1-1/\beta)\gamma'|\mathbf{n}|-\mathbf{n}^T\boldsymbol{\gamma}'}\\ &\le M^{-\beta}\sum_{\mathbf{n}:\,\boldsymbol{\gamma}^T\mathbf{n}>1}\phi^{|\mathbf{n}|}\\ &\le M^{-\beta}\left(\left(\sum_{n_1,\ldots,n_\ell\ge 0}\phi^{|\mathbf{n}|}\right)-1\right)\\ &= M^{-\beta}\left(\left(\prod_{j=1}^{\ell}\sum_{n_j=0}^{\infty}\phi^{n_j}\right)-1\right)\\ &= M^{-\beta}\left(\frac{1}{(1-\phi)^\ell}-1\right)\\ &= M^{-\beta}, \end{aligned} \tag{B.9}$$

where the first inequality is from (B.7) and the second inequality follows from (B.8). The last equality is because $1-\phi = 2^{-1/\ell}$. This completes the proof of Lemma 1.

B.3. Proof of Lemma 2

The Bernstein inequality (see, e.g., (1.21) in appendix 1 of Anthony and Bartlett 2009) asserts that, for any $z\ge 0$, we have

$$\mathbb{P}(|Y-\mathbb{E}[Y]|>z)\le 2\exp\left(-\frac{z^2/2}{\sum_{i=1}^n\mathbb{E}[(X_i-\mathbb{E}[X_i])^2]+bz/3}\right). \tag{B.10}$$

We note that $\mathbb{E}[(X_i-\mathbb{E}[X_i])^2]\le\mathbb{E}[X_i^2]\le\mathbb{E}[X_i b]\le b\overline{\lambda}$. Plugging this into (B.10), we obtain

$$\mathbb{P}(|Y-\mathbb{E}[Y]|>z)\le 2\exp\left(-\frac{z^2/2}{nb\overline{\lambda}+bz/3}\right). \tag{B.11}$$

This implies (20) and completes the proof of Lemma 2.

B.4. Proof of Lemma 3

For every j, and according to our definition (5) of the tail exponent of an arrival process, there is a random variable $\overline{A}_j$ that dominates $A_j(t)$ for all $t \geq 0$ and for which all moments of order less than $1+\gamma_j$ are finite; see (3). Because $\gamma_j > 0$, we have $\int_0^\infty \mathbb{P}(\overline{A}_j > x)\,dx = \mathbb{E}[\overline{A}_j] < \infty$. Therefore, there exists a constant $M_2 > 0$ such that, for any $M > M_2$ and every j,

$$\int_{M/\eta \log M}^{\infty} \mathbb{P}(\overline{A}_j > x)\,dx \leq \frac{\gamma\epsilon}{60C\sqrt{\ell}}. \tag{B.12}$$

Note that the choice of M_2 is independent of T. For the rest of the proof, we fix M_2 and assume that $M > M_2$.

Recall that $\mathbf{A}^*(\tau) = \min\{\mathbf{A}(\tau), \theta_\tau\}$ so that

$$A_j(\tau) - A_j^*(\tau) = \max\{0, A_j(\tau) - \theta_\tau\}. \tag{B.13}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[A_j(\tau)] - \mathbb{E}[A_j^*(\tau)] &= \mathbb{E}[\max\{0, A_j(\tau) - \theta_\tau\}] \\
&= \int_0^\infty \mathbb{P}(\max\{0, A_j(\tau) - \theta_\tau\} > x)\,dx \\
&= \int_0^\infty \mathbb{P}(A_j(\tau) - \theta_\tau > x)\,dx \\
&= \int_{\theta_\tau}^\infty \mathbb{P}(A_j(\tau) > x)\,dx \\
&\leq \int_{M/\eta \log M}^\infty \mathbb{P}(A_j(\tau) > x)\,dx \\
&\leq \int_{M/\eta \log M}^\infty \mathbb{P}(\overline{A}_j > x)\,dx \\
&\leq \frac{\gamma\epsilon}{60C\sqrt{\ell}},
\end{aligned} \tag{B.14}$$

where the fourth equality uses a change of variables from $x + \theta_\tau$ to x, the first inequality uses the fact $\theta_\tau \leq M/\eta \log M$ for all $\tau \geq 0$ (see the definition (11) of θ_t), the second inequality is because $\overline{A}_j$ dominates $A_j(\tau)$, and the last inequality follows from (B.12). Therefore, for any $\tau \geq 0$,

$$\|\mathbb{E}[\mathbf{A}(\tau)] - \mathbb{E}[\mathbf{A}^*(\tau)]\| \leq \sqrt{\ell} \max_{j=1,\ldots,\ell} \left\{\mathbb{E}[A_j(\tau)] - \mathbb{E}[A_j^*(\tau)]\right\} \leq \frac{\gamma\epsilon}{60C}. \tag{B.15}$$

We now introduce some "simpler" events whose occurrence are shown to imply the event $\mathcal{E}^{\text{fluc}}(T, M)$ that is introduced in (21):

$$\mathcal{E}^* = \left\{ \left\| \sum_{\tau=t_0}^{t} (\mathbf{A}^*(\tau) - \mathbb{E}[\mathbf{A}^*(\tau)]) \right\| \leq \frac{\gamma\epsilon}{30C}(M + T - t_0), \quad \text{for } 0 \leq t_0 \leq t < T \right\}, \tag{B.16}$$

and for every j,

$$\mathcal{E}_j^* = \left\{ \left| \sum_{\tau=t_0}^{t} (A_j^*(\tau) - \mathbb{E}[A_j^*(\tau)]) \right| \leq \frac{\gamma\epsilon}{30C\ell}(M + T - t_0), \quad \text{for } 0 \leq t_0 \leq t < T \right\}, \tag{B.17}$$

and we observe that the events $\mathcal{E}_1^*, \ldots, \mathcal{E}_\ell^*$ imply the event $\mathcal{E}^*$. Then, the union bound, applied to the complements of these events, implies that

$$\mathbb{P}(\mathcal{E}^*) \geq 1 - \sum_{j=1}^{\ell} (1 - \mathbb{P}(\mathcal{E}_j^*)). \tag{B.18}$$

We now argue that the event $\mathcal{E}^*$ implies the event $\mathcal{E}^{\text{fluc}}(T, M)$ defined in (21). Indeed, suppose that event $\mathcal{E}^*$ occurs. Then, as long as $0 \leq t_0 \leq t < T$, we obtain

$$\begin{aligned}
&\left\| \sum_{\tau=t_0}^{t} (\mathbf{A}^*(\tau) - \boldsymbol{\lambda}^*) \right\| \\
&= \left\| \sum_{\tau=t_0}^{t} (\mathbf{A}^*(\tau) - \mathbb{E}[\mathbf{A}^*(\tau)]) + \sum_{\tau=t_0}^{t} (\mathbb{E}[\mathbf{A}^*(\tau)] - \mathbb{E}[\mathbf{A}(\tau)]) + \sum_{\tau=t_0}^{t} (\mathbb{E}[\mathbf{A}(\tau)] - \boldsymbol{\lambda}^*) \right\| \\
&\leq \left\| \sum_{\tau=t_0}^{t} (\mathbf{A}^*(\tau) - \mathbb{E}[\mathbf{A}^*(\tau)]) \right\| + \sum_{\tau=t_0}^{t} \|\mathbb{E}[\mathbf{A}^*(\tau)] - \mathbb{E}[\mathbf{A}(\tau)]\| + \sum_{\tau=t_0}^{t} \|\mathbb{E}[\mathbf{A}(\tau)] - \boldsymbol{\lambda}^*\| \\
&\leq \frac{\gamma\epsilon}{30C}(M + T - t_0) + \frac{\gamma\epsilon}{60C}(t - t_0) + \frac{\gamma\epsilon}{20C}(t - t_0) \\
&\leq \frac{\gamma\epsilon}{10C}(M + T - t_0),
\end{aligned} \tag{B.19}$$

where, in the second inequality, we use the definition of the event $\mathcal{E}^*$ to bound the first term, (B.15) to bound the second term, and (14) to bound the third term. The last inequality is because $t < T$. Thus, the event $\mathcal{E}^*$ implies the event $\mathcal{E}^{\text{fluc}}(T,M)$, and

$$\mathbb{P}(\mathcal{E}^{\text{fluc}}(T,M)) \geq \mathbb{P}(\mathcal{E}^*) \geq 1 - \sum_{j=1}^{\ell} (1 - \mathbb{P}(\mathcal{E}_j^*)), \tag{B.20}$$

where the second inequality follows from (B.18).

To complete the proof, we derive an upper bound for $1 - \mathbb{P}(\mathcal{E}_j^*)$. As a first step, we obtain a relation between various constants, which reflects the fact that η is chosen large enough. We have

$$\begin{aligned}
\frac{\gamma^2\epsilon^2\eta}{2(30C\ell)^2\overline{\mu} + 20\gamma\epsilon C\ell} &\geq \frac{\gamma^2\epsilon^2\eta}{1800C^2\ell^2\overline{\mu} + 20\overline{\mu}C\ell} \\
&\geq \frac{\gamma^2\epsilon^2}{1820C^2\ell^2\overline{\mu}} \cdot \eta \\
&= \frac{\gamma^2\epsilon^2}{1820C^2\ell^2\overline{\mu}} \cdot \frac{8000C^2\ell^2\overline{\mu}}{\gamma^2\epsilon^2} \\
&> 4,
\end{aligned} \tag{B.21}$$

where the first inequality is due to the assumptions $\gamma \leq 1$ and the fact $\epsilon < \overline{\mu}$, which is evident from the definition (16) of $\overline{\mu}$; the second inequality is because $C \geq 1$, and the equality follows from the definition of η in (17).

Finally, we note that, for any $\tau \geq 0$ and every j,

$$\mathbb{E}[A_j^*(\tau)] \leq \mathbb{E}[A_j(\tau)] \leq \|\mathbb{E}[\mathbf{A}(\tau)]\| \leq \|\boldsymbol{\lambda}^*\| + \epsilon \leq \overline{\mu}. \tag{B.22}$$

For any t_0 and t with $0 \leq t_0 \leq t < T$, using the fact that $A_j^*(\tau)$ is bounded above by θ_τ, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{\tau=t_0}^{t} (A_j^*(\tau) - \mathbb{E}[A_j^*(\tau)])\right| > \frac{\gamma\epsilon}{30C\ell}(M+T-t_0)\right) \\
&\quad\leq 2\exp\left(-\frac{\left(\frac{\gamma\epsilon}{30C\ell}(M+T-t_0)\right)^2}{2\frac{M+T-t_0}{\eta\log(M+T-t_0)} \cdot \left(\overline{\mu}(t-t_0+1) + \frac{\gamma\epsilon}{90C\ell}(M+T-t_0)\right)}\right) \\
&\quad\leq 2\exp\left(-\frac{\gamma^2\epsilon^2\eta\log(M+T-t_0)}{2(30C\ell)^2\overline{\mu} + 20\gamma\epsilon C\ell}\right) \\
&\quad\leq 2\exp(-4\log(M+t-t_0)) \\
&\quad= \frac{2}{(M+T-t_0)^4},
\end{aligned} \tag{B.23}$$

where the first inequality follows from Lemma 2 with $z = (\gamma\epsilon/30C\ell)(M+T-t_0)$, $b = (M+T-t_0)/\eta\log(M+T-t_0) = \theta_{t_0} \geq \theta_\tau$, $\overline{\lambda} = \overline{\mu}$ (see (B.22)), and $n = t - t_0 + 1$; the second inequality is because $t - t_0 + 1 < M + T - t_0$, and the third inequality is due to (B.21). Therefore, for every j,

$$\begin{aligned}
1 - \mathbb{P}(\mathcal{E}_j^*) &\leq \sum_{t_0=0}^{T-1}\sum_{t=t_0}^{T-1} \mathbb{P}\left(\left|\sum_{\tau=t_0}^{t} (A_j^*(\tau) - \mathbb{E}[A_j^*(\tau)])\right| > \frac{\gamma\epsilon}{30C\ell}(M+T-t_0)\right) \\
&\leq \sum_{t_0=0}^{T-1}\sum_{t=t_0}^{T-1} \frac{2}{(M+T-t_0)^4} \\
&= \sum_{t_0=0}^{T-1} \frac{2(T-t_0)}{(M+T-t_0)^4} \\
&\leq \sum_{t_0=0}^{T-1} \frac{2}{(M+T-t_0)^3} \\
&\leq \sum_{\tau=1}^{\infty} \frac{2}{(M+\tau)^3} \\
&\leq \int_0^{\infty} \frac{2}{(M+x)^3} dx \\
&= M^{-2},
\end{aligned} \tag{B.24}$$

where the first inequality is from the union bound and the second inequality is due to (B.23). Plugging (B.24) into (B.20), we obtain $\mathbb{P}(\mathcal{E}^{\text{fluc}}(T,M)) \geq 1 - \ell M^{-2}$. This completes the proof of Lemma 3.

B.5. Proof of Lemma 4

We start with the proof of the first part and assume that W is ϵ-invariant. We prove the result under the additional assumption that the set W is closed. This is without loss of generality for the following reason. Given an ϵ-invariant set W, let $\overline{W}$ be its closure. Because fluid trajectories under the MW policy are continuous functions of their initial conditions, it is not hard to see that $\overline{W}$ is also ϵ-invariant. Once we show the result for closed sets, we have established that $\overline{W}$ is ϵ-attracting. Finally, because $d(\mathbf{x}, W) = d(\mathbf{x}, \overline{W})$ for all $\mathbf{x}$, we can conclude that W is also ϵ-attracting.

Having assumed that W is closed, we now consider a fluid trajectory $\mathbf{q}(\cdot)$ initialized with $\mathbf{q}(0) = \mathbf{q}_0 \geq \mathbf{0}$ with $\mathbf{q}_0 \notin W$. Then, there exists some $\mathbf{x}_0 \in W$, which is closest to $\mathbf{q}_0$. Let $\mathbf{x}(\cdot)$ be a fluid trajectory initialized at $\mathbf{x}(0) = \mathbf{x}_0$ and corresponding to the rate vector

$$\boldsymbol{\lambda} = \boldsymbol{\lambda}^* + \epsilon \frac{\mathbf{q}_0 - \mathbf{x}_0}{\|\mathbf{q}_0 - \mathbf{x}_0\|}. \tag{B.25}$$

Because W is ϵ-invariant and $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\| = \epsilon$, we have $\mathbf{x}(t) \in W$ for all $t \geq 0$. In particular, $d(\mathbf{q}(t), W) \leq \|\mathbf{q}(t) - \mathbf{x}(t)\|$ for all $t \geq 0$. Furthermore, equality holds at time $t = 0$. This implies that

$$\frac{d}{dt} d(\mathbf{q}(t), W)\Big|_{t=0} \leq \frac{d}{dt} \|\mathbf{q}(t) - \mathbf{x}(t)\|\Big|_{t=0}. \tag{B.26}$$

From the fluid equations (see Definition 2), we have $\dot{\mathbf{x}}(0) \in \overline{\mathcal{D}}_{\boldsymbol{\lambda}}(\mathbf{x}_0) = \boldsymbol{\lambda} - \overline{\mathcal{M}}(\mathbf{x}_0)$. Equivalently, there exist coefficients $\alpha_{\boldsymbol{\mu}} \geq 0$ for $\boldsymbol{\mu} \in \mathcal{M}(\mathbf{x}_0)$ such that $\sum_{\boldsymbol{\mu} \in \mathcal{M}(\mathbf{x}_0)} \alpha_{\boldsymbol{\mu}} = 1$ and

$$\dot{\mathbf{x}}(0) = \boldsymbol{\lambda} - \sum_{\boldsymbol{\mu} \in \mathcal{M}(\mathbf{x}_0)} \alpha_{\boldsymbol{\mu}} \boldsymbol{\mu}. \tag{B.27}$$

Similarly, let $\beta_{\boldsymbol{\nu}} \geq 0$ for $\boldsymbol{\nu} \in \mathcal{M}(\mathbf{q}_0)$ be a set of coefficients such that $\sum_{\boldsymbol{\nu} \in \mathcal{M}(\mathbf{q}_0)} \beta_{\boldsymbol{\nu}} = 1$ and

$$\dot{\mathbf{q}}(0) = \boldsymbol{\lambda}^* - \sum_{\boldsymbol{\nu} \in \mathcal{M}(\mathbf{q}_0)} \beta_{\boldsymbol{\nu}} \boldsymbol{\nu}. \tag{B.28}$$

For any $\boldsymbol{\mu} \in \mathcal{M}(\mathbf{x}_0)$ and any $\boldsymbol{\nu} \in \mathcal{M}(\mathbf{q}_0)$ because, by definition, $\boldsymbol{\mu}$ is a maximizer of $\mathbf{u}^T \mathbf{x}_0$ over all $\mathbf{u} \in \mathcal{M}$, we have $(\boldsymbol{\mu} - \boldsymbol{\nu})^T \mathbf{x}_0 \geq 0$. Similarly, $(\boldsymbol{\nu} - \boldsymbol{\mu})^T \mathbf{q}_0 \geq 0$. Combining these two inequalities, we obtain

$$(\boldsymbol{\mu} - \boldsymbol{\nu})^T (\mathbf{x}_0 - \mathbf{q}_0) = (\boldsymbol{\mu} - \boldsymbol{\nu})^T \mathbf{x}_0 + (\boldsymbol{\nu} - \boldsymbol{\mu})^T \mathbf{q}_0 \geq 0, \qquad \forall \boldsymbol{\mu} \in \mathcal{M}(\mathbf{x}_0),\ \boldsymbol{\nu} \in \mathcal{M}(\mathbf{q}_0). \tag{B.29}$$

Because $\sum_{\boldsymbol{\mu} \in \mathcal{M}(\mathbf{x}_0)} \alpha_{\boldsymbol{\mu}} = \sum_{\boldsymbol{\nu} \in \mathcal{M}(\mathbf{q}_0)} \beta_{\boldsymbol{\nu}} = 1$, it follows that

$$\left(\sum_{\boldsymbol{\mu} \in \mathcal{M}(\mathbf{x}_0)} \alpha_{\boldsymbol{\mu}} \boldsymbol{\mu} - \sum_{\boldsymbol{\nu} \in \mathcal{M}(\mathbf{q}_0)} \beta_{\boldsymbol{\nu}} \boldsymbol{\nu} \right)^T (\mathbf{x}_0 - \mathbf{q}_0) = \sum_{\boldsymbol{\mu} \in \mathcal{M}(\mathbf{x}_0)} \sum_{\boldsymbol{\nu} \in \mathcal{M}(\mathbf{q}_0)} \alpha_{\boldsymbol{\mu}} \beta_{\boldsymbol{\nu}} (\boldsymbol{\mu} - \boldsymbol{\nu})^T (\mathbf{x}_0 - \mathbf{q}_0) \geq 0. \tag{B.30}$$

Using (B.26), we have

$$\begin{aligned}
\frac{d}{dt} d(\mathbf{q}(t), W)\Big|_{t=0} &\leq \frac{d}{dt} \|\mathbf{q}(t) - \mathbf{x}(t)\|\Big|_{t=0} \\
&= \frac{(\dot{\mathbf{x}}(0) - \dot{\mathbf{q}}(0))^T (\mathbf{x}_0 - \mathbf{q}_0)}{\|\mathbf{x}_0 - \mathbf{q}_0\|} \\
&= \frac{1}{\|\mathbf{x}_0 - \mathbf{q}_0\|} \left(\boldsymbol{\lambda} - \sum_{\boldsymbol{\mu} \in \mathcal{M}(\mathbf{x}_0)} \alpha_{\boldsymbol{\mu}} \boldsymbol{\mu} - \boldsymbol{\lambda}^* + \sum_{\boldsymbol{\nu} \in \mathcal{M}(\mathbf{q}_0)} \beta_{\boldsymbol{\nu}} \boldsymbol{\nu} \right)^T (\mathbf{x}_0 - \mathbf{q}_0) \\
&= \frac{(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*)^T (\mathbf{x}_0 - \mathbf{q}_0)}{\|\mathbf{x}_0 - \mathbf{q}_0\|} - \Big(\sum_{\boldsymbol{\mu} \in \mathcal{M}(\mathbf{x}_0)} \alpha_{\boldsymbol{\mu}} \boldsymbol{\mu} - \sum_{\boldsymbol{\nu} \in \mathcal{M}(\mathbf{q}_0)} \beta_{\boldsymbol{\nu}} \boldsymbol{\nu} \Big)^T \frac{\mathbf{x}_0 - \mathbf{q}_0}{\|\mathbf{x}_0 - \mathbf{q}_0\|} \\
&\leq \frac{(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*)^T (\mathbf{x}_0 - \mathbf{q}_0)}{\|\mathbf{x}_0 - \mathbf{q}_0\|} \\
&= \left(\frac{\epsilon (\mathbf{q}_0 - \mathbf{x}_0)}{\|\mathbf{x}_0 - \mathbf{q}_0\|} \right)^T \frac{(\mathbf{x}_0 - \mathbf{q}_0)}{\|\mathbf{x}_0 - \mathbf{q}_0\|} \\
&= -\epsilon.
\end{aligned} \tag{B.31}$$

Here, the second equality follows from (B.27) and (B.28), the second inequality is due to (B.30), and the fourth equality uses the definition of $\boldsymbol{\lambda}$ in (B.25). Thus, W is ϵ-attracting.

For the proof of the second part, we fix some $\mathbf{n}$ and consider the set $W(\mathbf{n})$ as in Definition 6. Let $\mathbf{x}$ be an element of $W(\mathbf{n})$. Then, there exists an ϵ-JF($\mathbf{n}$) trajectory $\mathbf{x}(\cdot)$ that reaches $\mathbf{x}$. Consider now a fluid trajectory $\mathbf{q}(\cdot)$, corresponding to $\boldsymbol{\lambda}$, for some $\boldsymbol{\lambda}$ with $\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^*\|\le\epsilon$ and initialized at $\mathbf{x}$. The concatenation of the trajectories $\mathbf{x}(\cdot)$ and $\mathbf{q}(\cdot)$ is an ϵ-JF($\mathbf{n}$) trajectory, and therefore, any point that it can reach also belongs to $W(\mathbf{n})$. Thus, $W(\mathbf{n})$ is ϵ-invariant as claimed.

B.6. Proof of Lemma 5

As in the statement of the lemma, we fix some $M\ge M_3$ and times that satisfy $0\le t_0<t_1\le T$. We also fix a sample path under which the event $\mathcal{E}^{\text{fluc}}(T,M)$ occurs, and the interval $(t_0,\,t_1)$ is jump-free.

Let $\mathbf{q}(\cdot)$ be a fluid trajectory corresponding to the arrival rate $\boldsymbol{\lambda}^*$ and initialized with $\mathbf{q}(t_0+1)=\mathbf{Q}(t_0+1)$. Then,

$$\begin{aligned}\|\mathbf{Q}(t_1)-\mathbf{q}(t_1)\| &\le C\left(1+\|\boldsymbol{\lambda}^*\|+\max_{t\in(t_0,t_1)}\left\|\sum_{\tau=t_0+1}^{t}(\mathbf{A}(\tau)-\boldsymbol{\lambda}^*)\right\|\right)\\ &= C\left(1+\|\boldsymbol{\lambda}^*\|+\max_{t\in(t_0,t_1)}\left\|\sum_{\tau=t_0+1}^{t}(\mathbf{A}^*(\tau)-\boldsymbol{\lambda}^*)\right\|\right)\\ &\le C\left(1+\|\boldsymbol{\lambda}^*\|+\frac{\gamma\epsilon}{10C}(M+T-t_0)\right)\\ &= \frac{\gamma\epsilon}{10}(M+T-t_0)+(\|\boldsymbol{\lambda}^*\|+1)C,\end{aligned}\tag{B.32}$$

where the first inequality follows from Theorem 4, the first equality is because $(t_0,\,t_1)$ is jump-free, and as a result, $\mathbf{A}^*(\tau)=\mathbf{A}(\tau)$ for all $\tau\in(t_0,t_1)$, and the second inequality is because of the occurrence of the event $\mathcal{E}^{\text{fluc}}(T,M)$.[8]

Moreover, from Lemma 4, $W(t_0+1)$ is ϵ-attracting. Furthermore, because $(t_0,\,t_1)$ is a jump-free interval, we have $W(t_1)=W(t_0+1)$. Therefore,

$$\begin{aligned}d(\mathbf{q}(t_1),\,W(t_1)) &= d(\mathbf{q}(t_1),\,W(t_0+1))\\ &\le \max\{0,\,d(\mathbf{q}(t_0+1),\,W(t_0+1))-(t_1-t_0-1)\epsilon\},\end{aligned}\tag{B.33}$$

where the inequality rests on (23).

We now consider the effect of possible jumps at time t_0. Let $k\ge0$ be the number of jumps that occur at time t_0 in different entries of $\mathbf{A}(t_0)$. Without loss of generality, suppose that $A_1,\dots,A_k$ undergo a jump at time t_0. Let $\mathbf{J}$ be an ℓ-dimensional vector with the first k entries equal to $A_1(t_0),\dots,A_k(t_0)$ and all other entries equal to zero. Then, $\mathbf{A}(t_0)-\mathbf{J}$ is an ℓ-dimensional vector whose first k entries are zero; all other entries are jump-free and are, therefore, bounded by $\theta_{t_0}=(M+T-t_0)/(\eta\log(M+T-t_0))$. Thus,

$$\|\mathbf{A}(t_0)-\mathbf{J}\|\le\frac{(M+T-t_0)\ell}{\eta\log(M+T-t_0)}.\tag{B.34}$$

A key consequence of our definition of the sets $W(t)$ is that if t_0 is a jump time, then some components of the vector $\mathbf{N}(t_0)$ are larger than those of $\mathbf{N}(t_0-1)$ so that the set $W(t_0+1)$ is larger than the set $W(t_0)$. In particular, if $\mathbf{x}\in W(t_0)$, then $\mathbf{x}+\mathbf{J}\in W(t_0+1)$. Now, let $\mathbf{x}$ be a point in the closure of $W(t_0)$ that is closest to $\mathbf{Q}(t_0)$, that is,

$$\mathbf{x}\in\operatorname*{argmin}_{\mathbf{y}\in\text{closure}(W(t_0))}\|\mathbf{y}-\mathbf{Q}(t_0)\|.$$

Because $\mathbf{x}\in\text{closure}(W(t_0))$, it follows that $\mathbf{x}+\mathbf{J}\in\text{closure}(W(t_0+1))$. Therefore,

$$\begin{aligned}d(\mathbf{q}(t_0+1),\,W(t_0+1)) &\le d(\mathbf{q}(t_0+1),\,\mathbf{x}+\mathbf{J})\\ &= d([\mathbf{Q}(t_0)-\boldsymbol{\mu}(t_0)]^+ +\mathbf{A}(t_0),\,\mathbf{x}+\mathbf{J})\\ &\le d([\mathbf{Q}(t_0)-\boldsymbol{\mu}(t_0)]^+ +\mathbf{A}(t_0),\,\mathbf{Q}(t_0)+\mathbf{A}(t_0))\\ &\quad + d(\mathbf{Q}(t_0)+\mathbf{A}(t_0),\,\mathbf{Q}(t_0)+\mathbf{J})+d(\mathbf{Q}(t_0)+\mathbf{J},\,\mathbf{x}+\mathbf{J})\\ &= d([\mathbf{Q}(t_0)-\boldsymbol{\mu}(t_0)]^+,\,\mathbf{Q}(t_0))+d(\mathbf{A}(t_0),\,\mathbf{J})+d(\mathbf{Q}(t_0),\,\mathbf{x})\\ &\le \|\boldsymbol{\mu}(t_0)\|+d(\mathbf{A}(t_0),\,\mathbf{J})+d(\mathbf{Q}(t_0),\,\mathbf{x})\\ &\le \overline{\mu}+\frac{(M+T-t_0)\ell}{\eta\log(M+T-t_0)}+d(\mathbf{Q}(t_0),\,\mathbf{x})\\ &= \overline{\mu}+\frac{(M+T-t_0)\ell}{\eta\log(M+T-t_0)}+d(\mathbf{Q}(t_0),\,W(t_0)),\end{aligned}\tag{B.35}$$

where the first inequality is because $\mathbf{x}+\mathbf{J}\in\text{closure}(W(t_0+1))$, the first equality is from the evolution formula for the MW dynamics in (1) and the initialization $\mathbf{q}(t_0+1)=\mathbf{Q}(t_0+1)$ for $\mathbf{q}(\cdot)$, the last inequality follows from (B.34), and the last equality is from the definition of $\mathbf{x}$.

Combining (B.32), (B.33), and (B.35), we obtain, for $M \geq M_3$,

$$\begin{aligned}
d(\mathbf{Q}(t_1), W(t_1)) &\leq d(\mathbf{Q}(t_1), \mathbf{q}(t_1)) + d(\mathbf{q}(t_1), W(t_1)) \\
&\leq \frac{\gamma\epsilon\,(M+T-t_0)}{10} + (\|\boldsymbol{\lambda}^*\|+1)C + d(\mathbf{q}(t_1), W(t_1)) \\
&\leq \frac{\gamma\epsilon\,(M+T-t_0)}{10} + (\|\boldsymbol{\lambda}^*\|+1)C + \max\{0,\, d(\mathbf{q}(t_0+1), W(t_0+1)) - (t_1-t_0-1)\epsilon\} \\
&\leq \frac{\gamma\epsilon\,(M+T-t_0)}{10} + (\|\boldsymbol{\lambda}^*\|+1)C \\
&\quad + \max\left\{0,\, \left[d(\mathbf{Q}(t_0), W(t_0)) + \overline{\mu} + \frac{(M+T-t_0)\ell}{\eta\log(M+T-t_0)}\right] - (t_1-t_0-1)\epsilon\right\} \\
&\leq \max\{0,\, d(\mathbf{Q}(t_0), W(t_0)) - (t_1-t_0-1)\epsilon\} \\
&\quad + \overline{\mu} + \frac{(M+T-t_0)\ell}{\eta\log(M+T-t_0)} + \frac{\gamma\epsilon\,(M+T-t_0)}{10} + (\|\boldsymbol{\lambda}^*\|+1)C \\
&\leq \max\{0,\, d(\mathbf{Q}(t_0), W(t_0)) - (t_1-t_0)\epsilon\} + \epsilon \\
&\quad + \frac{(M+T-t_0)\ell}{\eta\log(M+T-t_0)} + \frac{\gamma\epsilon\,(M+T-t_0)}{10} + (\|\boldsymbol{\lambda}^*\|+1)C + \overline{\mu} \\
&\leq \max\{0,\, d(\mathbf{Q}(t_0), W(t_0)) - (t_1-t_0)\epsilon\} + \frac{\gamma\epsilon\,(M+T-t_0)}{6},
\end{aligned} \tag{B.36}$$

where the second inequality is from (B.32), the third inequality is due to (B.33), the fourth inequality follows from (B.35), and the last inequality is because M has been chosen large enough as in (25). This completes the proof of Lemma 5.

B.7. Proof of Lemma 6

Let $N(t)$ be the cardinality of $\mathbf{N}(t)$, that is, the number of jumps up to time t. We also use the convention $N(-1) = 0$. When the sample path is such that the event $\mathcal{E}^{\text{jump}}(T,M)$ occurs, then $N(t) \leq 1/\gamma$ for all $t < T$. Thus, for any $t \in [0,T)$,

$$\frac{(N(t)+2)\,\gamma\epsilon}{3} \leq \frac{(1/\gamma+2)\,\gamma\epsilon}{3} = \frac{(1+2\gamma)\epsilon}{3} \leq \epsilon, \tag{B.37}$$

where we have used our assumption that $\gamma \leq 1$.

We fix a sample path of the arrival process $\mathbf{A}(\cdot)$ under which both events $\mathcal{E}^{\text{jump}}(T,M)$ and $\mathcal{E}^{\text{fluc}}(T,M)$ occur. We use strong induction to prove that, for any $t \in [0,T]$,

$$d(\mathbf{Q}(t), W(t)) \leq \frac{(N(t-1)+2)\,\gamma\epsilon}{6}\,(M+T-t). \tag{B.38}$$

To establish the base case of the induction, we use the assumption $\mathbf{Q}(0) = \mathbf{0}$ and the fact that $\mathbf{0} \in W(0)$ (because ϵ-JF trajectories are zero for negative times). Thus, $d(\mathbf{Q}(0), W(0)) = 0$, and therefore, (B.38) holds for $t = 0$.

For the induction step, we consider a time $t_1 \in (0,T]$ and assume that (B.38) holds for all $t < t_1$. We show that (B.38) also holds for time t_1. Let

$$t_0 = \max\{0, T - (2(T-t_1)+M)\} = \max\{0, 2t_1 - T - M\}. \tag{B.39}$$

Note that either $t_0 = 0 < t_1$ or $t_0 = 2t_1 - T - M \leq t_1 - M < t_1$ so that we always have $t_0 < t_1$. We consider two cases.

Case B.1. For the first case, we assume that the interval (t_0, t_1) is jump-free.[9] We consider two subcases. If $t_0 = 0$, then

$$\begin{aligned}
d(\mathbf{Q}(t_1), W(t_1)) &\leq \max\{0,\, d(\mathbf{Q}(t_0), W(t_0)) - (t_1-t_0)\epsilon\} + \frac{\epsilon\gamma}{6}\,(M+T-t_0) \\
&= \frac{\epsilon\gamma}{6}\,(M+T-t_0) \\
&\leq \frac{\epsilon\gamma}{6}\,(M+T-(2t_1-T-M)) \\
&= \frac{2\epsilon\gamma}{6}\,(M+T-t_1),
\end{aligned} \tag{B.40}$$

where the first inequality is due to Lemma 5 and the first equality is because of the assumptions $t_0 = 0$ and $\mathbf{Q}(0) = \mathbf{0}$ together with the observation that $\mathbf{0} \in W(0)$. The second inequality follows from (B.39), which implies that $0 = t_0 \geq 2t_1 - T - M$. In particular, (B.38) holds for $t = t_1$.

For the second subcase, we assume that $t_0 > 0$, in which case, $t_0 = 2t_1 - T - M$. Then,

$$\begin{aligned}
&d(\mathbf{Q}(t_1), W(t_1))\\
&\le \max\{0, d(\mathbf{Q}(t_0), W(t_0)) - (t_1 - t_0)\epsilon\} + \frac{\epsilon\gamma}{6}(M + T - t_0)\\
&\le \max\left\{0, \frac{(N(t_0 - 1) + 2)\gamma\epsilon}{6}(M + T - t_0) - (t_1 - t_0)\epsilon\right\} + \frac{\epsilon\gamma}{6}(M + T - t_0)\\
&= \max\left\{0, \frac{(N(t_0 - 1) + 2)\gamma\epsilon}{6}(2M + 2T - 2t_1) - (M + T - t_1)\epsilon\right\} + \frac{\epsilon\gamma}{6}(2M + 2T - 2t_1)\\
&= \max\left\{0, \left(\frac{(N(t_0 - 1) + 2)\gamma\epsilon}{3} - \epsilon\right)(M + T - t_1)\right\} + \frac{2\epsilon\gamma}{6}(M + T - t_1)\\
&= \frac{2\epsilon\gamma}{6}(M + T - t_1),
\end{aligned} \tag{B.41}$$

where the first inequality follows from Lemma 5, the second inequality is due to the induction hypothesis (B.38), the first equality uses the substitution $t_0 = 2t_1 - T - M$, and the last equality is due to (B.37). Thus, (B.38) again holds for $t = t_1$. This completes the induction step for Case B.1.

Case B.2. For the second case, we assume that there is a jump in the interval (t_0, t_1). Let $\hat{t}_0$ be the last jump time in the interval (t_0, t_1) so that $(\hat{t}_0, t_1)$ is jump-free. Because $\hat{t}_0$ is the last jump time, we have

$$N(t_1 - 1) = N(\hat{t}_0) \ge N(\hat{t}_0 - 1) + 1, \tag{B.42}$$

where the inequality is because there is at least one jump at $\hat{t}_0$ (we say "at least" because, at time $\hat{t}_0$, we could have jumps at multiple components of $\mathbf{A}(\cdot)$. Consequently,

$$\begin{aligned}
&d(\mathbf{Q}(t_1), W(t_1))\\
&\le \max\left\{0, d(\mathbf{Q}(\hat{t}_0), W(\hat{t}_0)) - (t_1 - \hat{t}_0)\epsilon\right\} + \frac{\epsilon\gamma}{6}(M + T - \hat{t}_0)\\
&\le \max\left\{0, \frac{(N(\hat{t}_0 - 1) + 2)\gamma\epsilon}{6}(M + T - \hat{t}_0) - (t_1 - \hat{t}_0)\epsilon\right\} + \frac{\epsilon\gamma}{6}(M + T - \hat{t}_0)\\
&= \max\left\{\frac{\epsilon\gamma}{6}(M + T - \hat{t}_0), \frac{(N(\hat{t}_0 - 1) + 3)\gamma\epsilon}{6}(M + T - \hat{t}_0) - (t_1 - \hat{t}_0)\epsilon\right\}\\
&\le \max\left(\frac{\epsilon\gamma}{6}(M + T - \hat{t}_0), \frac{(N(t_1 - 1) + 2)\gamma\epsilon}{6}(M + T - \hat{t}_0) - (t_1 - \hat{t}_0)\epsilon\right)\\
&\le \max\left\{\frac{\epsilon\gamma}{6}(2M + 2T - 2t_1), \frac{(N(t_1 - 1) + 2)\gamma\epsilon}{6}(M + T - t_1) + \left(\frac{(N(t_1 - 1) + 2)\gamma\epsilon}{6} - \epsilon\right)(t_1 - \hat{t}_0)\right\}\\
&\le \max\left\{\frac{2\epsilon\gamma}{6}(M + T - t_1), \frac{(N(t_1 - 1) + 2)\gamma\epsilon}{6}(M + T - t_1)\right\}\\
&= \frac{(N(t_1 - 1) + 2)\gamma\epsilon}{6}(M + T - t_1),
\end{aligned} \tag{B.43}$$

where the first inequality follows from Lemma 5 and the assumption that $(\hat{t}_0, t_1)$ is jump-free, the second inequality is due to the induction hypothesis (B.38), the third inequality is from (B.42), the fourth inequality is because $\hat{t}_0 \ge t_0 \ge 2t_1 - T - M$, and the last inequality is due to (B.37). Therefore, the induction step goes through for Case B.2 as well. This completes the proof of the induction and implies (B.38) for all $t \in [0, T]$.

Finally, letting $t = T$, (B.38) becomes

$$d(\mathbf{Q}(T), W(T)) \le \frac{(N(T - 1) + 2)\gamma\epsilon}{6} M \le \frac{M\epsilon}{2}, \tag{B.44}$$

where the last inequality is due to (B.37). This completes the proof of Lemma 6.

Appendix C. Proof of Lemmas for the Second Direction of Theorem 1: RDS ⇒ ϵ-JF Condition

C.1. Proof of Lemma 7

Suppose that the ϵ-JF($\boldsymbol{\gamma}$) condition fails to hold. Then, there exists a nonnegative integer vector $\mathbf{n}'$ with $\boldsymbol{\gamma}^T\mathbf{n}' \le 1$, an ϵ-JF($\mathbf{n}'$) trajectory $\mathbf{q}'(\cdot)$, and some time T such that $q'_m(T) > 0$. If $T = 0$, we can use right-continuity of trajectories to see that, without loss of generality, we can take T to be positive. We then define a new trajectory $\mathbf{q}(\cdot)$ by letting $\mathbf{q}(t) = \mathbf{q}'(tT)/T$ for all $t \ge 0$. It is not hard to verify that $\mathbf{q}(\cdot)$ is also an ϵ-JF($\mathbf{n}'$) trajectory, and satisfies $q_m(1) = q'_m(T)/T > 0$ so that the first property is satisfied.

Suppose now that some of the jumps of $\mathbf{q}(\cdot)$ happen after time 1. Let $\mathbf{n}$ be the vector that counts the number of jumps that take place until time 1. Starting with $\mathbf{q}(\cdot)$, we eliminate the jumps that happen after time 1 to obtain an ϵ-JF($\mathbf{n}$)

trajectory with $\boldsymbol{\gamma}^T\mathbf{n} \le \boldsymbol{\gamma}^T\mathbf{n}' \le 1$, all jumps in $[0,1]$, and $q_m(1) > 0$. By slightly perturbing the jump times and using a continuity argument, we can ensure that no two components have simultaneous jumps and also that all jump times belong to (0, 1) so that properties (b) and (c) in Lemma 7 are satisfied.

Finally, we can replace the arrival rates $\lambda_j(t)$ that drive the ϵ-JF trajectory by $\max\{\lambda_j(t),\epsilon'\}$, where $\epsilon' < \epsilon$ is a small positive constant. This ensures that $\inf_t \lambda_j(t) > 0$. Furthermore, using a continuity argument and as long as ϵ' is small enough, the property $q_m(1) > 0$ is preserved. This proves condition (d) in Lemma 7 and concludes the proof of the lemma.

C.2. Proof of Lemma 8

For simplicity, and without loss of generality, we present the proof for the case in which $t_0 = 0$. We let $\overline{T}_1$ be a large enough constant such that, for all $T \ge \overline{T}_1$ and all k, we have

$$(a_k - d)T \ge \max\{\overline{\mu}, \sqrt{T}\}. \tag{C.1}$$

This is possible because, according to the definition of d in (33), we have $\min_k a_k > d$.

We fix some $T \ge \overline{T}_1$ as well as some $k \in \{1,\dots,n\}$. We aim to show that the process has a substantial probability of a jump of size approximately a_{kT} during the interval $[\Theta_k T, (\Theta_k + d)T)$. Within the proof of this lemma, we use the symbol j (instead of j_k) to denote the index of the queue at which the kth jump took place.

From (C.1), we have $\log((a_k - d)T) \ge (1/2)\log T$. We then obtain, for any t, any j, and any $x \in [(a_k - d)T, a_k T]$,

$$\begin{aligned} f_{A_j(t)}(x) &= \frac{\overline{\lambda}_j(t)}{\sigma(\gamma_j)} \cdot x^{-(2+\gamma_j)} \log(x+1)\, \mathbb{1}(x \ge \overline{\mu}) \\ &= \frac{\overline{\lambda}_j(t)}{\sigma(\gamma_j)} \cdot x^{-(2+\gamma_j)} \log(x+1) \\ &\ge \frac{\overline{\lambda}_j(t)}{\sigma(\gamma_j)} \cdot (a_k T)^{-(2+\gamma_j)} \log((a_k - d)T) \\ &\ge \zeta T^{-(2+\gamma_j)} \log T, \end{aligned} \tag{C.2}$$

where the second equality is due to (C.1) and ζ is a positive constant chosen so that $\zeta \le \overline{\lambda}_i(t)/2\sigma(\gamma_i) a_k^{(2+\gamma_j)}$ for every i, k, and t. Note that ζ can be taken positive because, according to Lemma 7, we can assume that $\inf_t \overline{\lambda}_j(t) > 0$.

We define $\phi = \zeta d$. Then,

$$\begin{aligned} \mathbb{P}(A_j(t) \in [(a_k - d)T, a_k T]) &= \int_{(a_k - d)T}^{a_k T} f_{A_j(t)}(x)\, dx \\ &\ge (dT) \cdot \zeta T^{-(2+\gamma_j)} \log T \\ &= \phi\, T^{-(1+\gamma_j)} \log T. \end{aligned} \tag{C.3}$$

As in (39), but with $t_0 = 0$, we define

$$\mathbf{B}_k = \sum_{t=\Theta_k T}^{\Theta_k T + dT - 1} \mathbf{A}(t), \tag{C.4}$$

and for $t \in [\Theta_k T, (\Theta_k + d)T)$,

$$\mathbf{U}_t = \mathbf{B}_k - A_j(t)\mathbf{e}_j, \tag{C.5}$$

and note that $\|\mathbf{U}_t\| \le \|\mathbf{B}_k\|$ for every k and t. We have

$$\begin{aligned} \mathbb{P}(\|\mathbf{U}_t\| \ge 2dT\overline{\mu}) &\le \frac{\mathbb{E}[\|\mathbf{U}_t\|]}{2dT\overline{\mu}} \\ &\le \frac{\sum_{\tau=\Theta_k T}^{(\Theta_k + d)T - 1} \mathbb{E}[\|\mathbf{A}(\tau)\|]}{2dT\overline{\mu}} \\ &\le \frac{dT(\|\boldsymbol{\lambda}^*\| + \epsilon)}{2dT\overline{\mu}} \\ &\le \frac{1}{2}. \end{aligned} \tag{C.6}$$

where the first step makes use of the Markov inequality, and the last step uses the fact $\|\boldsymbol{\lambda}^*\| + \epsilon \le \overline{\mu}$.

For $t \in [\Theta_k T, (\Theta_k + d)T)$, let $\mathcal{E}_t$ be the event that $A_j(t) \in [(a_k - d)T, a_k T]$ and $\|\mathbf{U}_t\| \le 2dT\overline{\mu}$. Note that the term $A_j(t)$ is omitted from U_t, and therefore, $A_j(t)$ and U_t are independent. Thus, using (C.3) and (C.6), we obtain

$$\mathbb{P}(\mathcal{E}_t) \ge \frac{\phi}{2}\, T^{-(1+\gamma_j)} \log T, \quad \forall t \in [\Theta_k T, (\Theta_k + d)T). \tag{C.7}$$

In light of the definition of d in (33), we have $d(1+2\overline{\mu}) < a_k$. Therefore,

$$2dT\overline{\mu} < (a_k - d)T. \tag{C.8}$$

Thus, for any $t, \tau \in [\Theta_k T, (\Theta_k + d)T)$ with $\tau \neq t$, if $\|\mathbf{U}_t\| \le 2dT\overline{\mu}$, then

$$A_j(\tau) \le \|\mathbf{U}_t\| \le 2dT\overline{\mu} < (a_k - d)T. \tag{C.9}$$

Here, the first inequality follows because, if $\tau \neq t$, then $A_j(\tau)$ is one of the summands in the definition (C.5) of $\mathbf{U}_t$, and the last inequality is due to (C.8). Consequently, if $\|\mathbf{U}_t\| \le 2dT\overline{\mu}$, then $\mathcal{E}_\tau$ holds for no $\tau \in [\Theta_k T, (\Theta_k + d)T)$ with $\tau \neq t$, that is, for $t \neq \tau$, the events $\mathcal{E}_t$ and $\mathcal{E}_\tau$ are disjoint.

We now claim that, for any $t \in [\Theta_k T, \Theta_k T + dT)$, the event $\mathcal{E}_t$ implies the event $\mathcal{E}_k^{\text{jump}}$ defined in (40). Indeed, when event $\mathcal{E}_t$ occurs, we obtain

$$\begin{aligned}
\|\mathbf{B}_k - a_k T \boldsymbol{e}_j\| &= \|\mathbf{U}_t + A_j(t)\boldsymbol{e}_j - a_k T \boldsymbol{e}_j\| \\
&\le \|\mathbf{U}_t\| + \|A_j(t)\boldsymbol{e}_j - a_k T \boldsymbol{e}_j\| \\
&= \|\mathbf{U}_t\| + |A_j(t) - a_k T| \\
&\le 2dT\overline{\mu} + dT,
\end{aligned} \tag{C.10}$$

where the last inequality follows from the definition of $\mathcal{E}_t$. This shows that the event $\mathcal{E}_k^{\text{jump}}$ occurs as claimed. Let $\psi = \min\{1, d\phi/2\}$. We then have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_k^{\text{jump}}) &\ge \mathbb{P}\left(\bigcup_{t \in [\Theta_k T, \Theta_k T + dT)} \mathcal{E}_t \right) \\
&= \sum_{t=\Theta_k T}^{(\Theta_k + d)T - 1} \mathbb{P}(\mathcal{E}_t) \\
&\ge dT \cdot \frac{\phi T^{-(1+\gamma_j)} \log T}{2} \\
&\ge \psi T^{-\gamma_j} \log T,
\end{aligned} \tag{C.11}$$

where the first equality is because, for $t \neq \tau$, the events $\mathcal{E}_t$ and $\mathcal{E}_\tau$ are disjoint; the second inequality is due to (C.7), and the last one uses the definition of ψ. This completes the proof of Lemma 8.

C.3. Proof of Lemma 9

This proof is similar to the proof of Lemma 3 in Appendix B.4 although some of the details are different.

Recall that r is the number of piecewise constant pieces in the trajectory $\mathbf{q}^\epsilon(\cdot)$. We fix some $k \in \{0, 1, \ldots, n\}$ and let

$$\alpha = \frac{\gamma c}{32 C r}, \tag{C.12}$$

and

$$\eta = \frac{8\ell(6\ell\overline{\mu} + \alpha)}{3\alpha^2}. \tag{C.13}$$

Claim C.1. *There exists a $\overline{T}_2 \ge 8\ell$ such that, for any $T \ge \overline{T}_2$, any j, and any $t \in [\Theta_k T + dT, \Theta_{k+1} T)$, we have*

$$\int_{T/\eta \log T}^{\infty} \mathbb{P}(A_j(t) \ge x)\, dx \le \frac{\alpha}{2\sqrt{\ell}}, \tag{C.14}$$

and

$$\mathbb{P}\left(A_j(t) > \frac{T}{\eta \log T} \right) \le \frac{1}{4\ell T}. \tag{C.15}$$

Proof. Because all γ_j are positive, we can fix some δ such that $0 < \delta < \gamma_j$ for every j. Then, the density $f_{A_j(t)}(\cdot)$ in Definition 8 decays at least as fast as $x^{-(2+\delta)}$. More concretely, there exists a constant χ such that for all j and t, we have

$$f_{A_j(t)}(x) \le \chi x^{-(2+\delta)}, \qquad \forall x \ge \overline{\mu}.$$

We then have, for any time $t \in [\Theta_k T + dT, \Theta_{k+1} T)$, any $y \ge 1$, and any j,

$$\mathbb{P}(A_j(t) \ge y) \le \chi \int_y^\infty x^{-(2+\delta)}\, dx = \frac{\chi}{1+\delta} \cdot x^{-(1+\delta)}. \tag{C.16}$$

It then follows that, as T goes to infinity, both $T \cdot \mathbb{P}(A_j(\tau) > T/\eta \log T)$ and $\int_{T/\eta \log T}^\infty \mathbb{P}(A_j(\tau) \ge x)\, dx$ go to zero uniformly over all j and τ. Therefore, there exists a $\overline{T}_2$ such that, for any $T \ge \overline{T}_2$, (C.15) and (C.14) hold, and the claim follows. □

For the rest of the proof, we fix such a $\overline{T}_2$ and assume that $T \geq \overline{T}_2$. For any $\tau \in [\Theta_k T + dT, \Theta_{k+1} T)$ and every j, let

$$A_j^*(\tau) = \min\left\{A_j(\tau), \frac{T}{\eta \log T}\right\}. \tag{C.17}$$

We define the "no jumps" event $\mathcal{E}^=$ as follows:

$$\mathcal{E}^= = \{\mathbf{A}(\tau) = \mathbf{A}^*(\tau), \quad \text{for all} \tau \in [\Theta_k T + dT, \Theta_{k+1} T)\}. \tag{C.18}$$

Then,

$$\begin{aligned} 1 - \mathbb{P}(\mathcal{E}^=) &\leq \sum_{j=1}^{\ell} \sum_{\tau=\Theta_k T + dT}^{\Theta_{k+1} T} \mathbb{P}\left(A_j(\tau) > \frac{T}{\eta \log T}\right) \\ &\leq \sum_{j=1}^{\ell} \sum_{\tau=\Theta_k T + dT}^{\Theta_{k+1} T} \frac{1}{4\ell T} \\ &\leq \frac{1}{4}, \end{aligned} \tag{C.19}$$

where the first inequality uses the union bound, the second inequality follows from (C.15), and the last inequality uses the fact that d as defined in (33) is no larger than one. Moreover, for any j and any $\tau \in [\Theta_k T + dT, \Theta_{k+1} T)$ and using the same steps as in (B.14),

$$\mathbb{E}[A_j(\tau)] - \mathbb{E}[A_j^*(\tau)] \leq \int_{T/\eta \log T}^{\infty} \mathbb{P}(A_j(\tau) > x)\, dx \leq \frac{\alpha}{2\sqrt{\ell}},$$

where the last inequality follows from (C.14). Therefore, for any $\tau \geq 0$,

$$\|\mathbb{E}[\mathbf{A}(\tau)] - \mathbb{E}[\mathbf{A}^*(\tau)]\| \leq \sqrt{\ell} \max_j (\mathbb{E}[A_j(\tau)] - \mathbb{E}[A_j^*(\tau)]) \leq \frac{\alpha}{2}. \tag{C.20}$$

Consider now the following events:

$$\mathcal{E}^* = \left\{ \left\| \sum_{\tau=\Theta_k T + dT}^{t} (\mathbf{A}^*(\tau) - \mathbb{E}[\mathbf{A}^*(\tau)]) \right\| \leq \frac{\alpha T}{2}, \quad \forall t \in [\Theta_k T + dT, \Theta_{k+1} T) \right\}, \tag{C.21}$$

and for $j = 1, \ldots, \ell$,

$$\mathcal{E}_j^* = \left\{ \left| \sum_{\tau=\Theta_k T + dT}^{t} (A_j^*(\tau) - \mathbb{E}[A_j^*(\tau)]) \right| \leq \frac{\alpha T}{2\ell}, \quad \forall t \in [\Theta_k T + dT, \Theta_{k+1} T) \right\}. \tag{C.22}$$

Note that, if all of the events $\mathcal{E}_1^*, \ldots, \mathcal{E}_\ell^*$ occur, then $\mathcal{E}^*$ also occurs. By applying the union bound to the complement of these events, we have

$$1 - \mathbb{P}(\mathcal{E}^*) \leq \sum_{j=1}^{\ell} (1 - \mathbb{P}(\mathcal{E}_j^*)). \tag{C.23}$$

Consider a sample path under which the events $\mathcal{E}^=$ and $\mathcal{E}^*$ occur. Then, for any $t \in [\Theta_k T + dT, \Theta_{k+1} T)$,

$$\begin{aligned} \left\| \sum_{\tau=\Theta_k T + dT}^{t} (\mathbf{A}(\tau) - \mathbb{E}[\mathbf{A}(\tau)]) \right\| &= \left\| \sum_{\tau=\Theta_k T + dT}^{t} (\mathbf{A}^*(\tau) - \mathbb{E}[\mathbf{A}(\tau)]) \right\| \\ &= \left\| \sum_{\tau=\Theta_k T + dT}^{t} (\mathbf{A}^*(\tau) - \mathbb{E}[\mathbf{A}^*(\tau)]) + \sum_{\tau=\Theta_k T + dT}^{t} (\mathbb{E}[\mathbf{A}^*(\tau)] - \mathbb{E}[\mathbf{A}(\tau)]) \right\| \\ &\leq \left\| \sum_{\tau=\Theta_k T + dT}^{t} (\mathbf{A}^*(\tau) - \mathbb{E}[\mathbf{A}^*(\tau)]) \right\| + \sum_{\tau=\Theta_k T + dT}^{t} \|\mathbb{E}[\mathbf{A}^*(\tau)] - \mathbb{E}[\mathbf{A}(\tau)]\| \\ &\leq \left\| \sum_{\tau=\Theta_k T + dT}^{t} (\mathbf{A}^*(\tau) - \mathbb{E}[\mathbf{A}^*(\tau)]) \right\| + \frac{\alpha T}{2} \\ &\leq \frac{\alpha T}{2} + \frac{\alpha T}{2} \\ &= \frac{\gamma c T}{32 C r}, \end{aligned} \tag{C.24}$$

where the first equality is due to $\mathcal{E}^=$, the second inequality is due to (C.20), the third inequality follows from $\mathcal{E}^*$, and the last equality is from the definition of α in (C.12). Therefore, the events $\mathcal{E}^=$ and $\mathcal{E}^*$ imply the event $\mathcal{E}_k^{\text{fluc}}$. Using the union

bound on the complements of these events,

$$1-\mathbb{P}(\mathcal{E}_k^{\text{fluc}}) \le (1-\mathbb{P}(\mathcal{E}^{=})) + (1-\mathbb{P}(\mathcal{E}^{*})) \le \frac{1}{4} + \sum_{j=1}^{\ell}(1-\mathbb{P}(\mathcal{E}_j^{*})), \tag{C.25}$$

where the last inequality follows from (C.19) and (C.23). To complete the proof, we develop an upper bound on $1-\mathbb{P}(\mathcal{E}_j^{*})$.

For any $\tau \ge 0$ and every j and as in (B.22), we have

$$\mathbb{E}[A_j^{*}(\tau)] \le \mathbb{E}[A_j(\tau)] \le \|\mathbb{E}[\mathbf{A}(\tau)]\| \le \overline{\mu}. \tag{C.26}$$

For any $t \in [\Theta_k T + dT, \Theta_{k+1} T)$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{\tau=\Theta_k T+dT}^{t}(A_j^{*}(\tau)-\mathbb{E}[A_j^{*}(\tau)])\right| > \frac{\alpha T}{2\ell}\right)\\
&\quad\le 2\exp\left(-\frac{(\alpha T/2\ell)^2}{2\frac{T}{\eta\log T}\cdot(\overline{\mu}(t-\Theta_k T-dT+1)+\alpha T/6\ell)}\right)\\
&\quad\le 2\exp\left(-\frac{3\alpha^2\eta\log T}{4\ell(6\ell\overline{\mu}+\alpha)}\right)\\
&\quad= 2\exp(-2\log T)\\
&\quad= 2T^{-2},
\end{aligned} \tag{C.27}$$

where the first inequality follows from Lemma 2 with the identifications $z = \alpha T/2\ell$, $b = T/\eta\log T$, $\overline{\lambda} = \overline{\mu}$ (see (C.26)), and $n = t - \Theta_k T - dT + 1$; the second inequality is because $n \le T$; the equality follows from the definition of η in (C.13). Therefore, for every j,

$$\begin{aligned}
1-\mathbb{P}(\mathcal{E}_j^{*}) &\le \sum_{t=\Theta_k T+dT}^{T}\mathbb{P}\left(\left|\sum_{\tau=\Theta_k T+dT}^{t}(A_j^{*}(\tau)-\mathbb{E}[A_j^{*}(\tau)])\right| > \frac{\alpha T}{2\ell}\right)\\
&\le \sum_{t=1}^{T}\frac{2}{T^2}\\
&= 2T^{-1}\\
&\le \frac{1}{4\ell},
\end{aligned} \tag{C.28}$$

where the first inequality is from the union bound, the second inequality is due to (C.27), and the last inequality is from the condition $T \ge \overline{T}_2 \ge 8\ell$ in Claim C.1. Plugging (C.28) into (C.25), we obtain $\mathbb{P}(\mathcal{E}_k^{\text{fluc}}) \ge 1/2$. This completes the proof of Lemma 9.

C.4. Proof of Lemma 10

The proof of this lemma is essentially a continuity argument together with an induction that pieces together different time segments.

For simplicity and without loss of generality, we assume that the constant t_0 in the statement of the lemma is equal to zero. Note, however, that, with this convention $\mathbf{Q}(0)$ is, in general, nonzero.

For any $\boldsymbol{\lambda}, \mathbf{x} \in \mathbb{R}_+^{\ell}$, let

$$\boldsymbol{\xi}_{\boldsymbol{\lambda}}(\mathbf{x}) = \dot{\mathbf{q}}(0), \tag{C.29}$$

where $\mathbf{q}(\cdot)$ is the fluid trajectory corresponding to arrival rate $\boldsymbol{\lambda}$ and initialized with $\mathbf{q}(0) = \mathbf{x}$. In view of (9), we have $\boldsymbol{\xi}_{\boldsymbol{\lambda}}(\mathbf{x}) \in \overline{\mathcal{D}}_{\boldsymbol{\lambda}}(\mathbf{x})$. From the definition (8) of the set $\overline{\mathcal{D}}_{\boldsymbol{\lambda}}(\mathbf{x})$ of possible drifts and the standing assumption $\|\boldsymbol{\lambda}(\tau) - \boldsymbol{\lambda}^{*}\| \le \epsilon$ for all τ, we have

$$\|\boldsymbol{\xi}_{\boldsymbol{\lambda}(\tau)}(\mathbf{q}^{\epsilon}(\tau))\| \le \overline{\mu}, \qquad \forall \tau \ge 0, \tag{C.30}$$

where $\overline{\mu}$ is defined in (30). By taking into account the jump $a_k \boldsymbol{e}_{j_k}$ at time Θ_k and then integrating the drift $\boldsymbol{\xi}_{\boldsymbol{\lambda}(\tau)}$ over the jump-free interval $[\Theta_k, \Theta_k + d]$, we have

$$\|\mathbf{q}^{\epsilon}(\Theta_k + d) - \mathbf{q}^{\epsilon}(\Theta_k)\| \le d\overline{\mu}.$$

Noting also that $\mathbf{q}^{\epsilon}(\Theta_k) = \mathbf{q}^{\epsilon}(\Theta_k^{-}) + a_k \boldsymbol{e}_{j_k}$, we obtain

$$\|\mathbf{q}^{\epsilon}(\Theta_k + d) - \mathbf{q}^{\epsilon}(\Theta_k^{-}) - a_k \boldsymbol{e}_{j_k}\| \le d\overline{\mu}. \tag{C.31}$$

For $\tau \geq 0$, let $\overline{\boldsymbol{\mu}}^a(\tau) = \min\{\boldsymbol{\mu}(\tau), \mathbf{Q}(\tau)\}$, where the minimum is taken component-wise; thus, $\boldsymbol{\mu}^a(\tau)$ is the actual service received at time τ. It then follows from (1) that

$$\mathbf{Q}(\tau+1) = \mathbf{Q}(\tau) - \boldsymbol{\mu}^a(\tau) + \mathbf{A}(\tau). \tag{C.32}$$

We start by considering the intervals $[\Theta_k T, (\Theta_k + d)T]$ associated with jumps for $k = 1, \ldots, n$. We are working with a sample path for which the event $\mathcal{E}_k^{\text{jump}}$ occurs, Therefore,

$$\begin{aligned}
&\|\mathbf{Q}((\Theta_k + d)T) - (\mathbf{Q}(\Theta_k T) + Ta_k \boldsymbol{e}_{j_k})\| \\
&\quad = \left\| (\mathbf{Q}(\Theta_k T) + \sum_{\tau=\Theta_k T}^{\Theta_k T + dT - 1} (\mathbf{A}(\tau) - \boldsymbol{\mu}^a(\tau))) - (\mathbf{Q}(\Theta_k T) + Ta_k \boldsymbol{e}_{j_k}) \right\| \\
&\quad \leq \left\| \left(\sum_{\tau=\Theta_k T}^{\Theta_k T + dT - 1} \mathbf{A}(\tau) \right) - Ta_k \boldsymbol{e}_{j_k} \right\| + \left\| \sum_{\tau=\Theta_k T}^{\Theta_k T + dT - 1} \boldsymbol{\mu}^a(\tau) \right\| \\
&\quad \leq dT(1 + 2\overline{\mu}) + dT\overline{\mu} \\
&\quad = dT(1 + 3\overline{\mu}),
\end{aligned} \tag{C.33}$$

where the first equality is due to (C.32). The second inequality follows from the definition (40) of $\mathcal{E}_k^{\text{jump}}$ and the fact $\|\boldsymbol{\mu}^a(\tau)\| \leq \|\boldsymbol{\mu}(\tau)\| \leq \overline{\mu}$ for all $\tau \geq 0$.

Combining (C.33) with (C.31), it follows that, for $k = 1, \ldots, n$,

$$\begin{aligned}
&\|\mathbf{Q}((\Theta_k + d)T) - T\mathbf{q}^\epsilon(\Theta_k + d)\| \\
&\quad \leq \|\mathbf{Q}((\Theta_k + d)T) - (\mathbf{Q}(\Theta_k T) + Ta_k \boldsymbol{e}_{j_k})\| + \|(\mathbf{Q}(\Theta_k T) + Ta_k \boldsymbol{e}_{j_k}) - T(\mathbf{q}^\epsilon(\Theta_k^-) + a_k \boldsymbol{e}_{j_k})\| \\
&\qquad + \|T(\mathbf{q}^\epsilon(\Theta_k^-) + a_k \boldsymbol{e}_{j_k}) - T\mathbf{q}^\epsilon(\Theta_k + d)\| \\
&\quad \leq \|\mathbf{Q}((\Theta_k + d)T) - (\mathbf{Q}(\Theta_k T) + Ta_k \boldsymbol{e}_{j_k})\| + \|\mathbf{Q}(\Theta_k T) - T\mathbf{q}^\epsilon(\Theta_k^-)\| + dT\overline{\mu} \\
&\quad \leq dT(1 + 3\overline{\mu}) + \|\mathbf{Q}(\Theta_k T) - T\mathbf{q}^\epsilon(\Theta_k^-)\| + dT\overline{\mu} \\
&\quad = \|\mathbf{Q}(\Theta_k T) - T\mathbf{q}^\epsilon(\Theta_k^-)\| + dT(1 + 4\overline{\mu}),
\end{aligned}$$

where the second and third inequalities are due to (C.31) and (C.33), respectively. Using the definition (33) of d, we conclude that

$$\|\mathbf{Q}((\Theta_k + d)T) - T\mathbf{q}^\epsilon(\Theta_k + d)\| \leq \|\mathbf{Q}(\Theta_k T) - T\mathbf{q}^\epsilon(\Theta_k^-)\| + \frac{\gamma c T}{8}. \tag{C.34}$$

We have so far established that, if the two trajectories $\mathbf{Q}(\cdot)$ and $\mathbf{q}^\epsilon(\cdot)$ are close at the beginning of an interval $[\Theta_k T, (\Theta_k + d)T]$, they are also close at the end. We now need to establish a similar conclusion over intervals of the form $[(\Theta_k + d)T, \Theta_{k+1} T]$. We wish to capitalize on Theorem 4. That result, however, refers to fluid models with constant arrival rates. In contrast, our stochastic process has a piecewise constant arrival rate, and the same holds for the associated JF trajectory. We deal with this issue by applying Theorem 4 repeatedly for each one of the piecewise constant segments.

Let us fix some $k \in \{1, \ldots, n\}$ and recall that r is an upper bound on the number of subintervals in $[\Theta_k + d, \Theta_{k+1})$ during which $\boldsymbol{\lambda}(\cdot)$ stays constant. Let us then fix some times θ_i for $i = 1, \ldots, r+1$ such that

$$\Theta_k + d = \theta_1 < \cdots < \theta_r < \theta_{r+1} = \Theta_{k+1},$$

and such that $\boldsymbol{\lambda}(\cdot)$ is constant during each one of the intervals (θ_i, θ_{i+1}) for $i = 1, \ldots, r$.

Under our assumption that the sample path satisfies the event $\mathcal{E}_k^{\text{fluc}}$, we see that, during each one of the intervals $[\theta_i T, \theta_{i+1} T]$ and for $i = 1 \ldots, r$, we have

$$\begin{aligned}
\max_{t \in [\theta_i T, \theta_{i+1} T)} \left\| \sum_{\tau=\theta_i T}^{t} (\mathbf{A}(\tau) - \overline{\boldsymbol{\lambda}}(\tau)) \right\| &\leq 2 \max_{t \in [\Theta_k T + dT, \Theta_{k+1} T)} \left\| \sum_{\tau=\Theta_k T + dT}^{t} (\mathbf{A}(\tau) - \overline{\boldsymbol{\lambda}}(\tau)) \right\| \\
&\leq \frac{\gamma c T}{16 C r}.
\end{aligned} \tag{C.35}$$

We now note that the function $\tilde{\mathbf{q}}^\epsilon(\cdot)$ defined by $\tilde{\mathbf{q}}^\epsilon(t) = T\mathbf{q}^\epsilon(t/T)$ is also an ϵ-JF trajectory and, in particular, is a fluid trajectory during each interval $[\theta_i T, \theta_{i+1} T)$. We apply Theorem 4 over this interval. Using also the fact that $1 + \|\boldsymbol{\lambda}(\tau)\| \leq \overline{\mu}$, we obtain

$$\|\mathbf{Q}(\theta_{i+1} T) - T\mathbf{q}^\epsilon(\theta_{i+1}^-)\| \leq \|\mathbf{Q}(\theta_i T) - T\mathbf{q}^\epsilon(\theta_i^-)\| + C\overline{\mu} + C \cdot \frac{\gamma c T}{16 C r}. \tag{C.36}$$

By summing such inequalities, for $i = 1, \ldots, r$ and using the facts $\theta_1 = \Theta_k + d$, $\theta_{r+1} = \Theta_{k+1}$, and $\mathbf{q}^\epsilon((\Theta_k + d)^-) = \mathbf{q}^\epsilon(\Theta_k + d)$, we obtain

$$\|\mathbf{Q}(\Theta_{k+1}T) - T\mathbf{q}^\epsilon(\Theta_{k+1}^-)\| \le \|\mathbf{Q}((\Theta_k + d)T) - T\mathbf{q}^\epsilon(\Theta_k + d)\| + rC\overline{\mu} + \frac{\gamma cT}{16}. \tag{C.37}$$

We now add (C.34) and (C.37) to obtain

$$\|\mathbf{Q}(\Theta_{k+1}T) - T\mathbf{q}^\epsilon(\Theta_{k+1}^-)\| \le \|\mathbf{Q}(\Theta_k T) - T\mathbf{q}^\epsilon(\Theta_k^-)\| + rC\overline{\mu} + \frac{3\gamma cT}{16}.$$

We finally sum these bounds for $k = 1, \ldots, n$ and use the fact $\mathbf{q}^\epsilon(\Theta_{n+1}^-) = \mathbf{q}^\epsilon(1^-)$ to conclude that

$$\begin{aligned}\|\mathbf{Q}(T) - T\mathbf{q}^\epsilon(1^-)\| &\le \|\mathbf{Q}(\Theta_1 T) - T\mathbf{q}^\epsilon(\Theta_1^-)\| + nrC\overline{\mu} + \frac{3n\gamma cT}{16} \\ &\le \|\mathbf{Q}(\Theta_1 T) - T\mathbf{q}^\epsilon(\Theta_1^-)\| + nrC\overline{\mu} + \frac{3cT}{16},\end{aligned} \tag{C.38}$$

where we also make use of the property $n\gamma \le \boldsymbol{\gamma}^T\mathbf{n} \le 1$.

The interval $[0, \Theta_1)$ is to be treated a little differently as $\Theta_0 = 0$ is not a jump time. Even so, the argument leading to (C.37) applies verbatim and shows that

$$\|\mathbf{Q}(\Theta_1 T) - T\mathbf{q}^\epsilon(\Theta_1^-)\| \le \|\mathbf{Q}(0) - T\mathbf{q}^\epsilon(0)\| + rC\overline{\mu} + \frac{\gamma cT}{16} \le \|\mathbf{Q}(0)\| + rC\overline{\mu} + \frac{cT}{16},$$

where we make use of the fact that the fluid trajectory is initialized at zero and the inequality $\gamma \le 1$. Combining with (C.38), we finally conclude that

$$\|\mathbf{Q}(T) - T\mathbf{q}^\epsilon(1)\| \le \|\mathbf{Q}(0)\| + (n+1)rC\overline{\mu} + \frac{cT}{4}.$$

We now let $\overline{T}_3$ be large enough so that, for any $T \ge \overline{T}_3$, we have $(n+1)rC\overline{\mu} + (cT/4) + (cT/5) \le cT/2$. As long as $T \ge \overline{T}_3$ and using the assumption $\|\mathbf{Q}(0)\| \le cT/5$, we obtain

$$\|\mathbf{Q}(T) - T\mathbf{q}^\epsilon(1)\| \le \frac{cT}{5} + (n+1)rC\overline{\mu} + \frac{cT}{4} \le \frac{cT}{2}.$$

Finally, using $c = \mathbf{q}_m^\epsilon(1)$, we have

$$\begin{aligned}Q_m(T) &= T\mathbf{q}_m^\epsilon(1) - (T\mathbf{q}_m^\epsilon(1) - Q_m(T)) \\ &\ge T\mathbf{q}_m^\epsilon(1) - \|T\mathbf{q}^\epsilon(1) - \mathbf{Q}(T)\| \\ &\ge T\mathbf{q}_m^\epsilon(1) - \frac{cT}{2} \\ &= cT - \frac{cT}{2} \\ &= \frac{cT}{2},\end{aligned} \tag{C.39}$$

and this completes the proof of Lemma 10.

Endnotes

[1] This phenomenon can arise under other scheduling policies as well, for example, the generalized processor sharing policy (Borst et al. 2003).

[2] Strictly speaking, the results in Section 5 only establish the absence of robust delay stability, not delay instability for the specific arrival process distributions of our example. However, a slight modification of the proof in Section 8 shows that queue 3 is indeed delay unstable.

[3] Recall our assumption in Section 2 that, for any $\boldsymbol{\mu} \in \mathcal{M}$, the set $\mathcal{M}$ also contains all vectors that result from setting some entries of $\boldsymbol{\mu}$ equal to zero. It is shown in proposition 2 of Sharifnassab et al. (2020) that, under this assumption, the fluid model of Definition 2 is equivalent to the more standard, albeit more complicated, definitions of fluid models based on differential equations with boundary conditions.

[4] If $\gamma_j = \infty$ and $n_j = 0$, we use the convention $\infty \cdot 0 = 0$.

[5] We note that n may be equal to zero. For example, for a single unstable queue, an ϵ-JF trajectory becomes positive even in the absence of jumps. Such a system is not RDS, consistent with our result.

[6] To avoid notation clutter, we present the proof as if $\Theta_k T$ or $\Theta_k T + dT$ were integer, which is not necessarily the case. Everything goes through, with occasional trivial modifications, if a sum of the form $\sum_{t=a}^{b} c_t$ is interpreted as $\sum_{t=\lfloor a \rfloor}^{\lfloor b \rfloor} c_t$.

[7] The reason for the condition $\boldsymbol{\lambda}^* > \mathbf{0}$ is that, if $\boldsymbol{\lambda}^* = \mathbf{0}$, then a system is trivially 0-RDS, but a small perturbation that leads to positive arrival rates can result in a delay unstable system.

[8] In case $t_1 = t_0 + 1$, the interval (t_0, t_1) is empty, and the term involving a maximum over $t \in (t_0, t_1)$ is interpreted as zero.

[9] This includes the case in which $t_1 = t_0 + 1$ so that the interval (t_0, t_1) is empty.

References

Anantharam V (1989) How large delays build up in a GI/G/1 queue. *Queueing Systems* 5(4):345–367.
Anthony M, Bartlett PL (2009) *Neural Network Learning: Theoretical Foundations* (Cambridge University Press, New York).
Asmussen S (1996) Rare events in the presence of heavy tails. Glasserman P, Sigman K, Yao DD, eds. *Stochastic Networks* (Springer, New York), 197–214.
Bertsimas D, Gamarnik D, Tsitsiklis JN (2001) Performance of multiclass Markovian queueing networks via piecewise linear Lyapunov functions. *Ann. Appl. Probab.* 11(4):1384–1428.
Borst S, Boxma O, Jelenković P (2003) Reduced-load equivalence and induced burstiness in GPS queues with long-tailed traffic flows. *Queueing Systems* 43(4):273–306.
Chen B, Blanchet J, Rhee CH, Zwart B (2019) Efficient rare-event simulation for multiple jump events in regularly varying random walks and compound Poisson processes. *Math. Oper. Res.* 44(3):919–942.
Durrett R (1980) Conditioned limit theorems for random walks with negative drift. *Zeitschrift Wahrscheinlichkeitstheorie Verwandte Gebiete* 52(3):277–287.
Foss S, Korshunov D (2012) On large delays in multi-server queues with heavy tails. *Math. Oper. Res.* 37(2):201–218.
Foss S, Korshunov D, Zachary S (2011) *An Introduction to Heavy-Tailed and Subexponential Distributions* (Springer, New York).
Georgiadis L, Neely MJ, Tassiulas L (2006) *Resource Allocation and Cross-Layer Control in Wireless Networks* (Now Publishers Inc., Hanover, MA).
Jelenković P, Momčilović P (2003) Asymptotic loss probability in a finite buffer fluid queue with hetergeneous heavy-tailed on–off processes. *Ann. Appl. Probab.* 13(2):576–603.
Maguluri ST, Burle SK, Srikant R (2016) Optimal heavy-traffic queue length scaling in an incompletely saturated switch. *ACM SIGMETRICS Performance Evaluation Rev.* 44(1):13–24.
Markakis MG (2013) Scheduling in switched queueing networks with heavy-tailed traffic. Unpublished PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.
Markakis MG, Modiano E, Tsitsiklis JN (2014) Max-weight scheduling in queueing networks with heavy-tailed traffic. *IEEE/ACM Trans. Networking* 22(1):257–270.
Markakis MG, Modiano E, Tsitsiklis JN (2018) Delay analysis of the max-weight policy under heavy-tailed traffic via fluid approximations. *Math. Oper. Res.* 43(2):460–493.
Nair J, Jagannathan K, Wierman A (2015) When heavy-tailed and light-tailed flows compete: The response time tail under generalized max-weight scheduling. *IEEE/ACM Trans. Networking* 24(2):982–995.
Nair J, Wierman A, Zwart B (2020) *The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation* (Cambridge University Press, Cambridge, UK).
Neely MJ (2010) *Stochastic Network Optimization with Application to Communication and Queueing Systems.* Ying L, ed., Synthesis Lectures on Communication Networks and Algorithms 3.1 (Springer Nature, Cham, Switzerland).
Pakes A (1975) On the tails of waiting-time distributions. *J. Appl. Probab.* 12(3):555–564.
Shah D, Wischik D (2012) Switched networks with maximum weight policies: Fluid approximation and multiplicative state space collapse. *Ann. Appl. Probab.* 22(1):70–127.
Sharifnassab A, Tsitsiklis JN, Golestani SJ (2019) Sensitivity to cumulative perturbations for a class of piecewise constant hybrid systems. *IEEE Trans. Automatic Control* 65(3):1057–1072.
Sharifnassab A, Tsitsiklis JN, Golestani SJ (2020) Fluctuation bounds for the max-weight policy with applications to state space collapse. *Stochastic Systems* 10(3):223–250.
Tassiulas L, Ephremides A (1992) Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Automatic Control* 37(12):1936–1948.
Veraverbeke N (1977) Asymptotic behaviour of Wiener-Hopf factors of a random walk. *Stochastic Processes Appl.* 5(1):27–37.
Zwart B, Borst S, Mandjes M (2004) Exact asymptotics for fluid queues fed by multiple heavy-tailed on-off flows. *Ann. Appl. Probab.* 14(2):903–957.