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

STOCHASTIC SYSTEMS
Vol. 10, No. 3, September 2020, pp. 223–250
ISSN (print), ISSN 1946-5238 (online)

Fluctuation Bounds for the Max-Weight Policy with Applications to State Space Collapse

Arsalan Sharifnassab,[a] John N. Tsitsiklis,[b] S. Jamaloddin Golestani[a]

[a] Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran, 11365-11155; [b] Laboratory for Information and Decision Systems, Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Contact: asharif@mit.edu, https://orcid.org/0000-0002-3910-2878 (AS); jnt@mit.edu, https://orcid.org/0000-0003-2658-8239 (JNT); golestani@sharif.edu, https://orcid.org/0000-0001-9797-0748 (SJG)

Received: October 22, 2018
Revised: April 14, 2019
Accepted: April 23, 2019
Published Online in Articles in Advance: April 9, 2020

https://doi.org/10.1287/stsy.2019.0038

Abstract. We consider a multihop switched network operating under a max-weight scheduling policy and show that the distance between the queue length process and a fluid solution remains bounded by a constant multiple of the deviation of the cumulative arrival process from its average. We then exploit this result to prove matching upper and lower bounds for the time scale over which additive state space collapse (SSC) takes place. This implies, as two special cases, an additive SSC result in diffusion scaling under non-Markovian arrivals and, for the case of independent and identically distributed arrivals, an additive SSC result over an exponential time scale.

Keywords: max-weight scheduling policy • state space collapse • fluid model • sensitivity to cumulative perturbations

1. Introduction

The subject of this paper is a new line of analysis of the maximum-weight (MW) scheduling policy for single-hop and multihop networks. The main ingredient is a purely deterministic qualitative property of the queue dynamics: the trajectory followed by the queue vector under an MW policy tracks the trajectory of an associated deterministic fluid model within a constant multiple of the cumulative fluctuation of the arrival processes. With this property at hand, it is then a conceptually simple matter to translate concentration properties of the arrival processes to concentration properties for the queue vector. As a consequence, we can obtain the following:

a. New, simple derivations of existing results on the convergence to a fluid solution/trajectory and on state space collapse (SSC).

b. Stronger versions of existing SSC results, involving more general arrival processes and tighter concentration bounds.

c. An approach to obtaining new results that would seem rather difficult to establish with existing methods.

The core of our approach is the trajectory tracking result mentioned. The latter is, in turn, an adaptation of a similar result established in Sharifnassab et al. (2020) for a general class of continuous-time hybrid systems that move along the subdifferential of a piecewise linear convex potential function with finitely many pieces; other than an additional restriction to the positive orthant, a continuous-time variant of the MW dynamics turns out to be exactly of this type. However, a fair amount of additional work is needed to translate the general result to the standard, discrete-time, MW setting; see Theorem 2 and its proof.

1.1. General Background

We consider a multihop switched network with fixed routing, such as those arising in wireless networks (Lin et al. 2006) or switch fabrics (Ji et al. 2009). The network operates in discrete time and is driven by jobs (or packets) that arrive according to a stochastic, deterministic, or adversarial process. There is a scheduler that, at each time step, selects one of finitely many possible service vectors. These service vectors can be fairly arbitrary, reflecting interdependence constraints between different servers, for example, interference constraints in the context of wireless networks.

We focus on the popular MW scheduling policy (Tassiulas and Ephremides 1992), which operates as follows. At any time step, an MW policy associates to each queue a weight proportional to its length and

selects a service vector that maximizes the total weighted service. MW policies are known to have a number of attractive properties, such as maximal throughput (Tassiulas and Ephremides 1992, Eryilmaz et al. 2005, Georgiadis et al. 2006). In addition, under certain conditions, for example, a resource pooling assumption, they minimize the workload in the heavy traffic regime (Stolyar 2004). On the other hand, the queue size dynamics under MW policies are quite complex, and a detailed analysis is difficult.

A common way of reducing the complexity of the analysis involves a fluid approximation, also known as a fluid model. The fluid model relies on two simplifications that lead to a description in terms of a set of differential equations (see Section 2.3): (a) the dynamics evolve in continuous—rather than discrete—time and (b) the arrival process is replaced by a constant flow with the same average. The fluid model underlies a general technique for dealing with discrete-time networks: approximate the queue lengths by fluid solutions and then analyze the fluid model. This approach has proved useful in the study of the MW dynamics, leading to results on stability (Dai and Prabhakar 2000, Andrews et al. 2004), SSC (Stolyar 2004; Dai and Lin 2005; Shah and Wischik 2006, 2012), and delay stability under heavy-tailed arrivals (Markakis et al. 2016, 2018). A key ingredient behind such results is an understanding of the accuracy with which fluid solutions approximate the original queue length processes; this paper contributes to this understanding.

1.2. State Space Collapse Literature

A prominent application of fluid models is in establishing SSC, that is, that, in the heavy traffic regime, the queue length process stays close to a low-dimensional set for a long time and with high probability.[1]

Seminal SSC results for communication networks were given in the works of Reiman (1984), Bramson (1998), and Williams (1998). Subsequently, several works (Stolyar 2004; Shah and Wischik 2006, 2012; Kang et al. 2009) followed the general framework of Bramson (1998) to prove SSC under different scheduling policies, including for the case of MW policies. The general approach involves splitting an $O(r^2)$-long interval into intervals of length $O(r)$ and then showing that the fluid-scaled processes (i.e., $\widehat{q}(t) = \frac{1}{r}Q(\lfloor rt \rfloor)$) stay close to the fluid solutions in each one of these smaller intervals. The SSC results then follow from the property that the fluid solutions are attracted to a low-dimensional set, called the set of invariant points.

For single-hop networks with Markovian arrivals operating under a generalization of the MW policy, SSC was proved in Stolyar (2004). It was also shown, in Stolyar (2004), as a consequence of SSC, that the *workload process* converges to a reflected Brownian motion and that every MW-α policy[2] with $\alpha > 0$ minimizes this workload among all scheduling algorithms. The results of Stolyar (2004) were extended to multihop networks in Dai and Lin (2008) and to another generalization of MW policies in Shakkottai et al. (2004). For multihop networks with non-Markovian arrivals operating under MW-α, SSC under diffusion scaling was studied in Shah and Wischik (2012). Several works (Shah et al. 2010, Kang and Williams 2012) then used the results of Shah and Wischik (2012) to provide diffusion approximations for the MW dynamics. Finally, SSC has also facilitated the study of the steady-state expectation of the number of jobs in a network (Eryilmaz and Srikant 2012; Maguluri et al. 2014, 2016; Xie and Lu 2015; Maguluri and Srikant 2016; Wang et al. 2018).

1.3. Preview of Results

Our approach to the analysis of MW policies relies on a bound on the distance of the queue length processes from the fluid solutions in terms of the fluctuations of the cumulative arrival processes. In more detail, we consider a queue length process $Q(\cdot)$ driven by an arrival process $A(\cdot)$ with average rate λ and compare $Q(\cdot)$ with a fluid solution $q(\cdot)$ driven by a steady arrival stream with the same rate λ under the same initial conditions $q(0) = Q(0)$.

We already know that, under suitable scaling, the trajectories of the original discrete-time process remain close to the fluid solutions. Furthermore, the fluid model is well known to be nonexpansive[3] (Subramanian 2010). By combining these facts, it is quite plausible that one should be able to derive bounds of the form

$$\big\|Q(t) - q(t)\big\| \le c + \sum_{\tau=0}^{t-1} \big\|A(\tau) - \lambda\big\|, \tag{1}$$

where $A(t)$ is the vector of arrivals at each one of the queues at time t and c is a constant that is independent of $A(\cdot)$. However, our goal is to derive a stronger bound of the form

$$\big\|Q(t) - q(t)\big\| \le c + C \max_{k<t} \left\| \sum_{\tau=0}^{k} (A(\tau) - \lambda) \right\|, \tag{2}$$

for some constants c and C, independent of $A(\cdot)$ and λ. The bounds in Equations (1) and (2) are qualitatively different. Under common probabilistic assumptions and with high probability, $\sum_{\tau=0}^{t-1} \|A(\tau) - \lambda\|$ grows at a rate of t, whereas $\max_{k<t} \|\sum_{\tau=0}^{k}(A(\tau) - \lambda)\|$ only grows as (roughly) $\sqrt{t}$.

The sensitivity bound (2) allows us to make several contributions to the study of the MW policy.

a. We obtain a very simple proof of the convergence of fluid-scaled processes to fluid solutions; see Corollary 1.

b. We establish a strong SSC result for the MW policy. In particular, we derive an upper bound and a matching lower bound on the time scale over which additive SSC takes place; see Theorem 3. As a corollary, when the arrivals are independent and identically distributed (i.i.d.), we establish SSC for the process $\widetilde{q}(t) = Q(\lfloor e^{\alpha r} t \rfloor)/r$ for some constant α, that is, over an exponentially long time scale; see Corollary 2.

c. In another corollary, we establish an additive SSC result in diffusion scaling and under non-Markovian arrivals, which strengthens the currently available diffusion scaling results under the MW policy in several respects; see Section 2.4 for more details.

d. As is reported elsewhere, the sensitivity bound (2) provides tools that allow us to resolve an open problem from Markakis et al. (2018), on the delay stability in the presence of heavy-tailed traffic.

On the technical side, the proof of the sensitivity bound (2) exploits a similar bound from our earlier work (Sharifnassab et al. 2020) on the sensitivity of a class of hybrid subgradient dynamical systems to fluctuations of external inputs or disturbances. The main challenges here concern the transition from discrete to continuous time as well as the presence of boundary conditions as queue sizes are naturally constrained to be nonnegative. For the proof of our SSC results, we follow the general framework of Bramson (1998) while also taking advantage of the sensitivity bound (2). We believe that our tight characterization of the time scale over which SSC holds would have been very difficult without the strong sensitivity bound (2).

1.4. Outline

The rest of the paper is organized as follows. In the next section, we describe the network model and our conventions along with some background on fluid models and SSC. In Section 3, we present our central result, which is an inequality of the form (2); see Theorem 2. Then, in Section 4, we present our SSC results. We provide the proofs of our results in Sections 5 and 6 while relegating some of the details to the appendices for improved readability. Finally, in Section 7, we offer some concluding remarks and discuss possible extensions.

2. System Model and Preliminaries

In this section, we list our notational conventions, define the network model that we study, and go over the necessary background on fluid models and SSC.

2.1. Notation and Conventions

We denote by $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}_{++}$, $\mathbb{Z}$, $\mathbb{Z}_+$, and $\mathbb{N}$ the sets of real numbers, nonnegative reals, positive reals, integers, nonnegative integers, and positive integers, respectively.

A vector $v \in \mathbb{R}^n$, is always treated as a column vector with components v_i for $i = 1, \ldots, n$. We use v^T and $\|v\|_2$ to denote the transpose and the Euclidean norm, respectively, of v. For any two vectors v and u in $\mathbb{R}^n$, the relation $v \preceq u$ indicates that $v_i \le u_i$ for all i. Furthermore, we use $\min(v, u)$ to denote the componentwise minimum, that is, the vector with components $\min(v_i, u_i)$. For a vector $\mu \in \mathbb{R}^n$ and a set J of indices, we use $\sigma_{-J}(\mu)$ to denote the vector whose ith entry is equal to the ith entry of μ if $i \notin J$ and is equal to zero if $i \in J$. Finally, we let $\mathbb{1}_n$ be the n-dimensional vector with all components equal to one.

The notation $\mathrm{Conv}(\cdot)$ stands for the convex hull of a set of vectors in $\mathbb{R}^n$. Given a vector $v \in \mathbb{R}^n$ and a set $A \subseteq \mathbb{R}^n$, we let $v + A = \{v + x : x \in A\}$. We use $d(v, A)$ to denote the Euclidean distance of v from the set A. Furthermore, if W is an $n \times n$ matrix, we let $WA = \{Wx \mid x \in A\}$ be the image of the set A under the linear transformation associated with W. Given a vector $v \in \mathbb{R}^n$, $\mathrm{diag}(v)$ denotes the $n \times n$ diagonal matrix with the entries of v on its main diagonal.

Finally, for a function $f : \mathbb{R} \to \mathbb{R}$ and with a slight departure from standard conventions, we use either $\dot{f}(t)$ or $d^+f(t)/dt$ to denote the right derivative of f at t, assuming that it exists.

2.2. Network Model and the MW Policy

A discrete-time multihop network with fixed deterministic routing is specified by n queues, a nonnegative $n \times n$ routing matrix R, and a finite set $\mathcal{S} \subset \mathbb{R}^n_+$ of *actions* (or service vectors) that correspond to the different schedules that can be applied at any time.

The input to a network is a collection of n discrete-time, nonnegative arrival processes, described by functions $A_i : \mathbb{Z}_+ \to \mathbb{R}_+$, where $A_i(t)$ stands for the workload that arrives to queue i during the tth time slot. Whenever the arrival processes are ergodic stochastic processes, we define the *arrival rate* vector $\lambda \in \mathbb{R}^n_+$ as the vector whose ith component is the average of the process $A_i(\cdot)$. We use $Q_i(t)$ to denote the (always

nonnegative) workload at queue i at time t and $Q(t)$ to denote the corresponding workload vector. In the sequel, we use the terms *workload*, *queue size*, and *queue length* interchangeably. The evolution of $Q(t)$ is determined by the particular policy used to operate the network.

Given a network and an arrival process $A(\cdot)$, the evolution of the queue lengths is given by

$$Q(t+1) = Q(t) + A(t) + (R - I)\min\big(\mu(t), Q(t)\big), \quad \forall\, t \in \mathbb{Z}_+, \tag{3}$$

where $\mu(t)$ is the service vector chosen by the policy at time t, and as mentioned earlier, $\min\big(\mu(t), Q(t)\big)$ is to be interpreted componentwise. Equation (3) corresponds to the situation in which a time slot begins with a queue vector $Q(t)$, and then a service vector $\mu(t)$ is chosen and applied. Finally, the new arrivals $A(t)$ are recorded at the end of the time slot and contribute to the new queue vector $Q(t+1)$.

Note that the routing matrix R is deterministic and prespecified and is not affected by the queue sizes or the scheduling policy. Single-hop networks correspond to the special case in which R is the zero matrix. More generally, the most common case (single-path routing) is one in which the routing matrix has entries in $\{0,1\}$ with at most one nonzero entry in each column and in which the ijth entry being one indicates that any work completed at queue j is transferred to queue i for further processing. However, we allow for more general nonnegative matrices R because this additional freedom does not affect the main proofs and also allows for a simpler treatment of weighted MW policies; see Lemma 4 in the proof of Theorem 2.

The following assumption is in effect throughout the paper and is naturally valid in typical application contexts.

Assumption 1. *For any $\mu \in \mathcal{S}$ and any $i \in \{1, \ldots, n\}$, the set $\mathcal{S}$ also contains the vector $\sigma_{-\{i\}}(\mu)$, that is, the vector obtained by setting the ith component of μ to zero.*

According to Assumption 1, if a certain service vector μ is allowed, it is also possible to follow μ at all queues other than queue i while providing no service to queue i. In particular, the zero vector is always an element of $\mathcal{S}$. On the technical side, Assumption 1 appears innocuous; however, it is indispensable for the proof technique used in this paper and has also been made in earlier works (see section 4.1 of Stolyar (2005) and assumption 2.3 of Shah and Wischik (2012)).

We now proceed to define weighted max-weight (WMW) policies, which can be viewed as either a generalization of MW policies or as a special case of the broader class of MW-f policies[4] considered in Eryilmaz et al. (2005) and back-pressure–based utility maximization algorithms[5] considered in Stolyar (2005), Georgiadis et al. (2006), and Neely (2010). We are given a multihop network with n queues, as described, along with a positive vector $w \in \mathbb{R}^n_{++}$ and the associated diagonal matrix $W = \mathrm{diag}(w)$. For any $Q \in \mathbb{R}^n_+$, we let $\mathcal{S}_w(Q)$ be the set of maximizers of $Q^T W (I - R)\mu$:

$$\mathcal{S}_w(Q) \triangleq \operatorname*{argmax}_{\mu \in \mathcal{S}} Q^T W (I - R)\mu. \tag{4}$$

A WMW policy associated with w (or w-WMW, for short) chooses, at each time t, an arbitrary service vector $\mu(t) \in \mathcal{S}_w(Q(t))$.[6] An MW policy is a special case of a WMW policy, in which $w = \mathbb{1}_n$. When dealing with MW policies, we drop the subscript w and write $\mathcal{S}(Q)$ instead of $\mathcal{S}_{\mathbb{1}_n}(Q)$.

Consider an ergodic and Markovian arrival process with arrival rate vector $\lambda \in \mathbb{R}^n_+$ for which there exists some scheduling policy that stabilizes the network, that is, results in a positive recurrent process. The closure of the set of all such vectors λ is called the *capacity region* and is denoted by $\mathcal{C}$.

We now record a fact that is used later in the proofs of Lemma 5 and Claim 4. Fix some $\lambda \in \mathcal{C}$ in the capacity region and consider a stabilizing policy. We define f_i as the average departure rate from queue i. Then, the flow conservation property $f = \lambda + Rf$ implies that $\lambda = (I - R)f$. Moreover, following an argument similar to the one in section 3.C of Tassiulas and Ephremides (1992), there exists a vector c such that $f \preceq c \in \mathrm{Conv}(\mathcal{S})$. Assumption 1 then implies that $f \in \mathrm{Conv}(\mathcal{S})$, and as a result, $\lambda \in (I - R)\mathrm{Conv}(\mathcal{S})$. In conclusion,

$$\mathcal{C} \subseteq (I - R)\mathrm{Conv}(\mathcal{S}). \tag{5}$$

A remarkable property of MW and WMW policies is that they are *throughput optimal* in the sense that, for any λ in the interior of $\mathcal{C}$ and any ergodic Markovian arrival process with average arrival rate vector λ, the resulting process is positive recurrent (Tassiulas and Ephremides 1992). Similar throughput optimality results are available for extensions of MW, for example, for the so-called f-MW policies (Eryilmaz et al. 2005).

2.3. Fluid Model

The fluid model associated with the MW policy is a deterministic dynamical system that runs in continuous time and in which the arrival stream is replaced by a steady "fluid" arrival stream with rate vector λ. We are working with the following definition of the fluid model; somewhat different but equivalent definitions can be found in Shah and Wischik (2012) and Markakis et al. (2018).

Definition 1 (Fluid Solutions)**.** We are given an arrival rate vector λ and an initial queue length vector $q(0) \in \mathbb{R}_+^n$. A fluid model solution (or, simply, *fluid solution*) is an absolutely continuous function $q : \mathbb{R}_+ \to \mathbb{R}_+^n$ that, together with a collection of functions $s_\mu : \mathbb{R}_+ \to [0,1]$ for $\mu \in \mathcal{S}$ and another function $y : \mathbb{R}_+ \to \mathbb{R}_+^n$, satisfies the following relations almost everywhere:

$$\dot{q}(t) = \lambda + (R - I)\left(\sum_{\mu \in \mathcal{S}} s_\mu(t)\mu \; - \; y(t)\right), \tag{6}$$

$$\sum_{\mu \in \mathcal{S}} s_\mu(t) = 1, \tag{7}$$

$$y_i(t) \le \sum_{\mu \in \mathcal{S}} s_\mu(t)\mu_i, \qquad i = 1, \ldots, n, \tag{8}$$

$$\text{if } q_i(t) > 0, \quad \text{then } y_i(t) = 0, \qquad i = 1, \ldots, n, \tag{9}$$

$$\text{if } \mu \notin \mathcal{S}_w\big(q(t)\big), \quad \text{then } s_\mu(t) = 0, \qquad \forall\ \mu \in \mathcal{S}. \tag{10}$$

It is known that for any multihop network and any initial condition, a fluid solution always exists (see appendix A of Dai and Lin (2005) and lemma 9 of Markakis et al. (2018)) and is unique (see lemma 10 of Markakis et al. (2018)) even though the corresponding $s_\mu(\cdot)$ and $y(\cdot)$ need not be unique. Moreover, for $q(0) \ge 0$, (6)–(10) imply that $q(t)$ remains nonnegative for all subsequent times t. Later, in Proposition 2, we show that fluid solutions admit an alternative description as the trajectories of a related subgradient dynamical system.[7]

We are particularly interested in the set of invariant states of the fluid model, which, for any λ in the capacity region, is defined by (see theorem 5.4(iv) of Shah and Wischik (2012))

$$\mathcal{I}(\lambda) \triangleq \big\{q_0 \in \mathbb{R}_+^n \,\big|\, q(t) = q_0,\ \forall t, \text{ is a fluid solution}\big\}. \tag{11}$$

Our notation is chosen to emphasize the dependence on λ of the set of invariant states. We note that, if λ belongs to the interior of the capacity region, then $\mathcal{I}(\lambda)$ is a singleton equal to $\{0\}$. Thus, $\mathcal{I}(\lambda)$ can be nontrivial only if λ lies on the boundary of $\mathcal{C}$.

We now record a scaling property of the set of fluid solutions.

Lemma 1. *Consider a fluid solution $q(\cdot)$ and a constant $r > 0$. Let $\widehat{q}(t) = q(rt)/r$ for all $t \ge 0$. Then, $\widehat{q}(\cdot)$ is also a fluid solution.*

Proof. Note that the set of maximizing schedules in Equation (4) does not change when we scale the queue vector by a positive constant. Therefore, for any $t \ge 0$,

$$\mathcal{S}\big(\widehat{q}(t)\big) = \mathcal{S}\big(q(rt)/r\big) = \mathcal{S}\big(q(rt)\big). \tag{12}$$

Consider the functions $y(\cdot)$ and $s_\mu(\cdot)$ that, together with $q(\cdot)$, satisfy the fluid model relations (6)–(10). Let $\widehat{y}(t) = y(rt)$ and $\widehat{s}_\mu(t) = s_\mu(rt)$ for all $t \ge 0$ and all $\mu \in \mathcal{S}$. Then, it is easy to verify that $\widehat{q}(\cdot)$, $\widehat{y}(\cdot)$, and $\widehat{s}_\mu(\cdot)$ also satisfy (6)–(10). Therefore, $\widehat{q}(\cdot)$ is also a fluid solution. □

Suppose that λ is in the capacity region and that $q_0 \in \mathcal{I}(\lambda)$ so that $q(t) = q_0$ is a fluid solution. Then, Lemma 1 implies that, for any scalar $r > 0$, $q(t) = q_0/r$ is also a fluid solution, and therefore, $q_0/r \in \mathcal{I}(\lambda)$. Furthermore, it is not hard to see that the identically zero function is also a fluid solution so that $0 \in \mathcal{I}(\lambda)$. We conclude that $\mathcal{I}(\lambda)$ is a cone, that is,

$$\alpha\mathcal{I}(\lambda) = \mathcal{I}(\lambda), \qquad \forall\ \alpha > 0. \tag{13}$$

The interest in fluid solutions stems from the fact that they provide approximations to suitably scaled versions (i.e., under "fluid scaling") of the original process. We summarize here one such result, which is a special case of theorem 4.3 in Shah and Wischik (2012); similar results are given in Dai and Lin (2005), lemmas 4 and 5.

Proposition 1. *Fix some $\lambda \in \mathbb{R}_+^n$, $T > 0$, and $q_0 \in \mathbb{R}_+^n$. Letting r range over the positive integers, consider a sequence of arrival processes $A^r(\cdot)$ that satisfies*

$$\frac{1}{r}\max_{t\le rT}\left\|\sum_{\tau=0}^{t}(A^r(\tau)-\lambda)\right\| \underset{r\to\infty}{\longrightarrow} 0, \tag{14}$$

almost surely. Let $Q^r(\cdot)$ be the process generated according to Equation (3) when the arrival process is $A^r(\cdot)$ and the initial condition is $Q^r(0) = rq_0$. We define the continuous-time scaled processes $\widehat{q}^r(t) = Q^r(\lfloor rt\rfloor)/r$ and note that $\widehat{q}^r(0) = q_0$ for all r. Finally, let $q(\cdot)$ be a fluid solution under that particular vector λ, initialized with $q(0) = q_0$. Then,

$$\sup_{t\le T}\left\|\widehat{q}^r(t)-q(t)\right\| \underset{r\to\infty}{\longrightarrow} 0, \tag{15}$$

almost surely.

Condition (14) is typically satisfied under common probabilistic assumptions, for example, when $A^r(\cdot)$ is an i.i.d. process with mean λ and bounded domain or more generally of exponential type. Thus, loosely speaking, convergence of the arrival processes leads to convergence of the queue processes.

As we see in Section 3, our results allow for stronger statements; namely we show that the rate of convergence in Equation (14) provides bounds on the rate of convergence to the fluid solution in Equation (15); see Corollary 1.

2.4. State Space Collapse

In this section, we discuss known results about SSC under an MW policy, thus setting the stage for a comparison with the results we present in Section 4.

We consider the heavy traffic regime, in which the arrival rate vector gets arbitrarily close to some point λ on the outer boundary of the capacity region. In this regime, the average queue lengths typically tend to infinity, yet it is often the case that the queue length vector stays close to the set of invariant states, $\mathcal{I}(\lambda)$. This phenomenon is called SSC and has been studied extensively, mostly under the so-called diffusion scaling. In this scaling, we start with a sequence $Q^r(\cdot)$ of stochastic processes, indexed by $r \in \mathbb{N}$, and then proceed to study a sequence of scaled processes $\widehat{q}^r(\cdot)$, referred to as diffusion-scaled processes, defined by

$$\widehat{q}^r(t) = \frac{1}{r}Q^r(\lfloor r^2t\rfloor), \qquad t \ge 0. \tag{16}$$

The extent to which the queue length process stays close to the set of invariant states is, in general, determined by the magnitude of the fluctuations of the arrival process. It is, therefore, natural to start the analysis with some assumptions on these fluctuations. General SSC results, under the MW policy and some of its extensions, were provided in Shah and Wischik (2012) under the following assumption.[8]

Assumption 2 (Shah and Wischik (2012), assumption 2.5)**.** *Let $A^r(\cdot)$ be a sequence of arrival processes indexed by $r \in \mathbb{N}$. We assume that, for each r, $A^r(\cdot)$ is stationary[9] with mean λ^r and that $\lambda^r \to \lambda$ as $r \to \infty$. We, furthermore, assume that there exists a sequence $\delta_r \in \mathbb{R}_+$ converging to zero as $r \to \infty$ such that*

$$r\cdot\log^2 r\cdot \mathrm{P}\left(\max_{t\le r}\frac{1}{r}\left\|\sum_{\tau=0}^{t}(A^z(\tau)-\lambda^z)\right\|\ge\delta_r\right)\underset{r\to\infty}{\longrightarrow}0, \qquad \textit{uniformly in } z. \tag{17}$$

Note that Assumption 2 is quite general, not requiring the arrival processes to be i.i.d. or Markovian. Theorem 7.1 of Shah and Wischik (2012), slightly rephrased,[10] establishes that, for a network operating under an MW-f policy (a generalization of WMW policies and under certain conditions on f) for any $T > 0$ and under Assumption 2, the diffusion-scaled queue length processes $\widehat{q}^r(\cdot)$ satisfy, for any $\delta > 0$,

$$\mathrm{P}\left(\frac{\sup_{t\in[0,T]} d\left(\widehat{q}^r(t),\,\mathcal{I}(\lambda)\right)}{\max\left(1,\,\sup_{t\in[0,T]}\widehat{q}^r(t)\right)}>\delta\right)\underset{r\to\infty}{\longrightarrow}0, \tag{18}$$

when $\lim_{r\to\infty}\widehat{q}^r(0) = q_0$ for some $q_0 \in \mathcal{I}(\lambda)$.

The bound in (18) is referred to as *multiplicative* SSC. Yet there is a stronger notion, called *additive* SSC, which involves a bound similar to (18) but with the term $\max(\widehat{q}^r(t), 1)$ absent from the denumerator and which is known to hold under i.i.d. arrivals.

Theorem 1 (Shah et al. (2010), theorem 7.7)**.** *Consider a network operating under an MW-α policy with $\alpha \geq 1$ with i.i.d. and uniformly bounded arrivals with rate $\lambda^r \to \lambda$ for some $\lambda \in \mathcal{C}$ and the associated diffusion-scaled queue length processes $\widehat{q}^r(\cdot)$. Assume that $\lim_{r\to\infty} \widehat{q}^r(0) = q_0$ for some $q_0 \in \mathcal{I}(\lambda)$. Then,*[11] *for any $\delta > 0$,*

$$\mathrm{P}\left(\sup_{t\in[0,T]} d\big(\widehat{q}^r(t)\,,\, \mathcal{I}(\lambda)\big) > \delta \right) \underset{r\to\infty}{\longrightarrow} 0. \tag{19}$$

Compared with the literature, our results only apply to the case in which $\alpha = 1$ (i.e., the MW policy) but allow for queue-dependent weights so that the weight of queue i is $w_i Q_i$. More crucially, our results (see Section 4 and Theorem 3, in particular)

a. remain valid as long as $\lim_{r\to\infty} d(\widehat{q}^r(0), \mathcal{I}(\lambda)) = 0$, which is a weaker condition than $\lim_{r\to\infty} \widehat{q}^r(0) = q_0$, for a fixed $q_0 \in \mathcal{I}(\lambda)$;[12]

b. unlike Shah et al. (2010), we do not require the arrival process to be i.i.d. or bounded as long as the arrival process has certain concentration properties. Furthermore, the concentration properties that we require (see Definition 2) are weaker than Assumption 2 for the case of diffusion scaling (see Corollary 3); and

c. apply to scalings other than diffusion scaling and include a converse result that characterizes the possible scalings for which additive SSC holds.

We finally note another related line of work that studies a property similar to SSC, namely the extent to which the steady-state distribution is concentrated in a neighborhood of the set of invariant points. In particular, Maguluri and Srikant (2015) and Maguluri et al. (2016) have characterized the tail of the steady-state distribution of the distance from the set of invariant points for the case of an input-queued switch.

3. Main Result: Sensitivity

The backbone behind all of the results is the following main theorem.

Theorem 2 (Sensitivity of WMW Policy)**.** *For a network operating under a WMW policy, there exists a constant C, to be referred to as the sensitivity constant, that satisfies the following. Consider an arrival process $A(\cdot)$ and the corresponding queue length process $Q(\cdot)$. Let $q(\cdot)$ be a fluid solution corresponding to some $\lambda \geq 0$ and initialized with $q(0) = Q(0)$. Then, for any $k \in \mathbb{Z}_+$,*

$$\big\|Q(k) - q(k)\big\| \;\leq\; C\left(1 + \|\lambda\| + \max_{t<k} \left\| \sum_{\tau=0}^{t} (A(\tau) \,-\, \lambda) \right\| \right). \tag{20}$$

Note that the result holds without having to assume that λ lies inside the capacity region. The proof is given in Section 5, and the key steps are as follows. We show that the study of WMW policies can be reduced to the study of MW policies. Furthermore, given a network operating in discrete time under the MW policy, we introduce an associated continuous-time dynamical system, which we call the *induced* dynamical system. Next, we show that the fluid solutions and the queue length processes of the network can be viewed as unperturbed and perturbed trajectories of the induced dynamical system, respectively. We finally argue that the induced dynamical system falls within the class of subgradient systems that were studied in Sharifnassab et al. (2020) and apply the main result in that reference to prove (20). The reductions that are developed in the course of the proof may be of independent interest.

3.1. Convergence to Fluid Model Solutions

An immediate consequence of Theorem 2, together with Lemma 1, is a bound on the distance of the fluid-scaled process $\widehat{q}^r(t) = Q(\lfloor rt \rfloor)/r$ from a fluid solution $q(\cdot)$.

Corollary 1. *Consider a network operating under the WMW policy and let C be the constant in Theorem 2. Fix an arrival function $A(\cdot)$ and some $q_0 \in \mathbb{R}_+^n$. Let $Q^r(\cdot)$ be the process generated according to Equation (3) when the arrival process is $A(\cdot)$*

and the initial condition is $Q^r(0) = rq_0$. *Let* $\widehat{q}^r(t) = Q^r(\lfloor rt \rfloor)/r$. *Let* $q(\cdot)$ *be a fluid solution corresponding to some* $\lambda \in \mathbb{R}_+^n$ *and initialized at* $q(0) = q_0$. *Then, for any* $T > 0$,

$$\sup_{t \le T} \left\| \widehat{q}^r(t) - q(t) \right\| \le \frac{C}{r} \max_{t < rT} \left\| \sum_{\tau=0}^{t} (A(\tau) - \lambda) \right\| + O(1/r). \tag{21}$$

Corollary 1 strengthens (15) significantly. Any statistical assumptions on the fluctuations of the arrival process $A(\cdot)$ readily yield concrete upper bounds on the distance of the original process from its fluid counterpart.

4. State Space Collapse

In this section, we apply Theorem 2 to establish a general additive SSC result; see Theorem 3. We then continue with some corollaries on exponential scaling or diffusion scaling. Our approach can also be used to obtain results that apply in steady state. However, we do not go into that latter topic because such results can also be proved using simpler, more direct methods as in Maguluri and Srikant (2015) and Maguluri et al. (2016).

4.1. Definitions and Preliminaries

At the core of our proofs lies the following lemma, which asserts that fluid solutions are attracted to the set $\mathcal{I}(\lambda)$ of invariant states, which was defined in Equation (11). The proof of the lemma is given in Appendix A.

Lemma 2 (Attraction to the Set of Invariant States). *Consider a network operating under the MW policy and a vector* λ *in its capacity region. There exists a constant* $\alpha(\lambda) > 0$ *such that, for any fluid solution* $q(\cdot)$ *associated with* λ *and any time* t,

$$q(t) \notin \mathcal{I}(\lambda) \Longrightarrow \frac{d^+}{dt} d\big(q(t), \mathcal{I}(\lambda)\big) \le -\alpha(\lambda),$$

with this right-derivative being guaranteed to exist.

We continue with a definition that quantifies the rate at which a family of processes concentrates on its mean.

Definition 2 (f-Tailed Sequence of Random Processes). Consider a function $f : \mathbb{N} \times \mathbb{R}_+ \to \mathbb{R}+$ and a vector $\lambda \in \mathbb{R}^n$. Let $A^r(\cdot)$ be a sequence of random processes indexed by $r \in \mathbb{N}$. Assume that, for each r, $A^r(\cdot)$ is stationary and has expected value λ^r and that $\lim_{r\to\infty} \lambda^r = \lambda$. Suppose that, for every $\delta > 0$,

$$f(r,\delta)\, \mathrm{P}\left(\frac{1}{r} \sup_{t \le r} \left\| \sum_{\tau=0}^{t} (A^r(\tau) - \lambda^r) \right\| > \delta \right) \underset{r\to\infty}{\longrightarrow} 0. \tag{22}$$

Then, $A^r(\cdot)$ is said to be an f-tailed sequence of random processes with limit mean λ, and we refer to f as the *concentration rate function*.

Later, we show that the time scale over which SSC holds is almost proportional to the best possible concentration rate function f. We observe that any sequence of random processes that satisfies Assumption 2 is a sequence of f-tailed processes with $f(r,\delta) = r \cdot \log^2 r$. However, the reverse is not true: Assumption 2 involves an additional requirement of uniform convergence over all values of an additional indexing parameter z, whereas Definition 2 essentially only considers the case $z = r$. Thus, Definition 2 is less restrictive and easier to check and also seems more natural.

There are many processes whose concentration properties are well understood and that translate to the requirements in Definition 2 for a suitable concentration rate function f. We record one such fact in Lemma 3, which deals with bounded i.i.d. arrival processes and which is proved in Appendix B.

Lemma 3 (Bounded I.I.D. Processes are Exponential-Tailed). *Fix a vector* $\lambda \in \mathbb{R}_+^n$ *and a constant* $a > 0$. *Consider a sequence of random processes* $A^r(\cdot)$ *indexed by* $r \in \mathbb{N}$. *Suppose that, for every* r, *the random variables* $A^r(t)$ *are i.i.d. and that* $A^r(t) \in [0,a]^n$ *for all* t. *Denote the mean of* $A^r(t)$ *by* λ^r *and suppose that* $\lim_{r\to\infty} \lambda^r = \lambda$. *Take any constant* $\beta \in (0,2)$ *and let* $f(r,\delta) = \exp(\beta r\delta/na^2)$. *Then,* $A^r(\cdot)$ *is an* f*-tailed sequence of random processes with limit mean* λ.

Similar results are possible for arrival processes that are modulated by a finite and ergodic Markov chain. The boundedness assumption can also be removed under standard conditions on the moment-generating function of $A^r(t)$.

We now define processes involving a more general scaling of time as a generalization of the fluid and diffusion-scaled processes.

Definition 3 (g-Time-Scaled Processes)**.** Consider an increasing function $g : \mathbb{N} \to \mathbb{R}_+$ and a sequence $Q^r(\cdot)$ of random processes. Then, the corresponding sequence of g-time-scaled processes $\widehat{q}^r(\cdot)$ is defined as

$$\widehat{q}^r(t) = \frac{1}{r} Q^r\big(\lfloor g(r)t \rfloor\big) \tag{23}$$

for all $r \in \mathbb{N}$ and all $t \in \mathbb{R}_+$.

The fluid scaling and the diffusion scaling of a random process are particular g-time-scaled processes corresponding to $g(r) = r$ and $g(r) = r^2$, respectively. Definition 3 allows for a more general scaling of time.

4.2. Main SSC Result

We now present our main SSC result.

Theorem 3 (Strong State Space Collapse)**.** *Consider a network operating under a WMW policy and a vector λ in its capacity region with a corresponding set of invariant states $\mathcal{I}(\lambda)$. Fix some $T \in \mathbb{R}_+$ and let $\{\lambda^r\}$ be a sequence that converges to λ. Consider two functions $f : \mathbb{N} \times \mathbb{R}_+ \to \mathbb{R}_+$ and $g : \mathbb{N} \to \mathbb{R}_+$ with $\liminf_{r\to\infty} g(r)/r > 0$. Let $A^r(\cdot)$ be an f-tailed sequence of arrival processes with limit mean λ and let $\widehat{q}^r(\cdot)$ be a corresponding sequence of g-time-scaled queue length processes. Suppose that $d\big(\widehat{q}^r(0)\,,\,\mathcal{I}(\lambda)\big) \to 0$ as $r \to \infty$.*

a. *Suppose that, for every $\epsilon > 0$, we have $\liminf_{r\to\infty} rf(r,\epsilon)/g(r) > 0$. Then, for any $\delta > 0$,*

$$\mathrm{P}\left(\sup_{t\in[0,T]} d\big(\widehat{q}^r(t)\,,\,\mathcal{I}(\lambda)\big) > \delta\right) \underset{r\to\infty}{\longrightarrow} 0. \tag{24}$$

b. *Under the same assumptions as in part (a), we can also bound the rate of convergence in (24): for any $\delta > 0$, there exists an $\epsilon > 0$ such that*

$$\frac{rf(r,\epsilon)}{g(r)}\,\mathrm{P}\left(\sup_{t\in[0,T]} d\big(\widehat{q}^r(t)\,,\,\mathcal{I}(\lambda)\big) > \delta\right) \underset{r\to\infty}{\longrightarrow} 0. \tag{25}$$

Moreover, for the case of an MW policy, (25) holds for every $\epsilon \le \min(\delta, \alpha)/2C$, where C is the sensitivity constant of the network (see Theorem 2) and $\alpha = \alpha(\lambda)$ is the constant in Lemma 2.

c. *Conversely, suppose that $f : \mathbb{N} \times \mathbb{R}_+ \to \mathbb{R}_+$ and $g : \mathbb{N} \to \mathbb{R}_+$ are such that $\lim_{r\to\infty} rf(r,\epsilon)/g(r) = 0$ for every $\epsilon > 0$ and $\lim_{r\to\infty} g(r)/r = \infty$. Then, for any network operating under an MW policy, any arrival rate λ in its capacity region (excluding its extreme points), and any $q_0 \in \mathcal{I}(\lambda)$, there exists an f-tailed sequence of arrival processes satisfying (22) and a corresponding sequence of g-time-scaled processes $\widehat{q}^r(\cdot)$, $r \in \mathbb{N}$, initialized at $\widehat{q}^r(0) = q_0$ such that*

$$\mathrm{P}\left(\sup_{t\in[0,T]} d\big(\widehat{q}^r(t)\,,\,\mathcal{I}(\lambda)\big) > \delta\right) \underset{r\to\infty}{\longrightarrow} 1 \tag{26}$$

for all $\delta > 0$.

The proof of Theorem 3 is given in Section 6. Part (b) relies on a reduction of WMW dynamics to MW dynamics together with the facts that the queue length process stays close to a fluid solution (Theorem 2) and that a fluid solution is attracted to the invariant set $\mathcal{I}$ (Lemma 2). The proof of part (c) relies on an explicit construction.

We note that part (a) is a straightforward corollary of part (b). Nevertheless, we have included the statement of part (a) because it is in a form comparable to SSC results in the literature and also because it facilitates a comparison with the converse result in part (c).

Theorem 3 ties together the time scaling g over which SSC occurs and the concentration rate function, f, of the arrival processes. The underlying intuition is that, if the queue length process is initialized sufficiently close to $\mathcal{I}$, then it stays in an $r\delta$-neighborhood of $\mathcal{I}$ with high probability for a period of time proportional to $g(r)$. This enables us to prove additive SSC over time scales much longer than those underlying the diffusion scaling as in the next section.

4.3. Special Cases of SSC

In this section, we apply Theorem 3 to obtain more concrete SSC results. The first result concerns SSC over an exponentially large time scale. Although it refers to bounded i.i.d. processes, it admits straightforward

extensions to arrival processes with a concentration rate function f that grows exponentially with r as is the case whenever a suitable large deviations principle holds.

Corollary 2 (Bounded I.I.D. Arrivals: SSC over an Exponential Time Scale)**.** *Consider a network operating under an MW policy, a vector λ in its capacity region, a $\delta > 0$, and a sequence $A^r(\cdot)$ of arrival processes that satisfy the assumptions of Lemma 3. Consider a $\gamma < \min(\delta, \alpha)/(2Cna^2)$, where C is the input sensitivity constant of the network, $\alpha = \alpha(\lambda)$ is the constant in Lemma 2, and a is an upper bound on the size of arriving jobs (see Lemma 3). Consider the $e^{\gamma r}$-time scaling of the queue length processes,*

$$\widehat{q}^r(t) = \frac{1}{r} Q^r\big(\lfloor e^{\gamma r} t \rfloor\big), \tag{27}$$

and suppose that $d\big(\widehat{q}^r(0)\,,\, \mathcal{I}(\lambda)\big) \to 0$ as $r \to \infty$. Then, for any $T \in \mathbb{R}_+$,

$$e^{\gamma r}\, \mathrm{P}\left(\sup_{t\in[0,T]} d\big(\widehat{q}^r(t)\,,\, \mathcal{I}(\lambda)\big) > \delta \right) \underset{r\to\infty}{\longrightarrow} 0. \tag{28}$$

Proof. Let $\beta = \gamma/[\min(\delta, \alpha)/(2Cna^2)]$. Then, $\beta < 1$, and Lemma 3 implies that $A^r(\cdot)$ is an f-tailed sequence of processes for $f(r, \epsilon) = \exp\big(2\beta r\epsilon/na^2\big)$. Let $\epsilon = \min(\delta, \alpha)/2C$. Then, $\gamma = \beta\epsilon/na^2$. Let $g(r) = \exp(\gamma r) = \exp(\beta r\epsilon/na^2)$ be the time scaling in the definition (27) of $\widehat{q}^r(t)$. Then,

$$\frac{rf(r,\epsilon)}{g(r)} = \frac{r\exp\big(2\beta r\epsilon/na^2\big)}{\exp\big(\beta r\epsilon/na^2\big)} = r\exp\big(\beta r\epsilon/na^2\big) \;>\; \exp(\gamma r). \tag{29}$$

Therefore, the assumptions in part (b) of Theorem 3 are satisfied, and

$$\begin{aligned} &\limsup_{r\to\infty}\, e^{\gamma r}\, \mathrm{P}\left(\sup_{t\in[0,T]} d\big(\widehat{q}^r(t)\,,\, \mathcal{I}(\lambda)\big) > \delta \right) \\ &\qquad \le \limsup_{r\to\infty} \frac{rf(r,\epsilon)}{g(r)}\, \mathrm{P}\left(\sup_{t\in[0,T]} d\big(\widehat{q}^r(t)\,,\, \mathcal{I}(\lambda)\big) > \delta \right) = 0. \end{aligned} \tag{30}$$

Thus, (28) holds, which is the desired result. □

We note that part (c) of Theorem 3 provides a partial converse to Corollary 2: under i.i.d. arrivals with nonzero variance, additive SSC does not hold over a superexponential time scale.

The next corollary of Theorem 3(a) concerns additive SSC under diffusion scaling.

Corollary 3 (State Space Collapse in Diffusion Scaling)**.** *Consider a network operating under a WMW policy and a function $f : \mathbb{N} \times \mathbb{R}_+ \to \mathbb{R}_+$ such that $\liminf_{r\to\infty} f(r, \delta)/r > 0$ for all $\delta > 0$. Consider a λ in the capacity region, an f-tailed sequence of arrivals $A^r(\cdot)$ with limit mean λ, and a corresponding diffusion-scaled queue length processes $\widehat{q}^r(\cdot)$ (see (16)). Suppose that $d\big(\widehat{q}^r(0)\,,\, \mathcal{I}(\lambda)\big) \to 0$ as $r \to \infty$. Then, for any $T \in \mathbb{R}_+$,*

$$\mathrm{P}\left(\sup_{t\in[0,T]} d\big(\widehat{q}^r(t)\,,\, \mathcal{I}(\lambda)\big) > \delta \right) \underset{r\to\infty}{\longrightarrow} 0. \tag{31}$$

Corollary 3 strengthens Theorem 1 for the case of WMW policies in that the assumption of i.i.d. arrivals is removed. We only require a concentration property for the arrival process, such as

$$r\,\mathrm{P}\left(\sup_{t\le r} \frac{1}{r} \left\| \sum_{\tau=0}^{t} (A^r(\tau) - \lambda) \right\| \ge \delta \right) \underset{r\to\infty}{\longrightarrow} 0, \qquad \forall \delta > 0, \tag{32}$$

which is even weaker than Assumption 2. Moreover, under an MW policy and i.i.d. arrivals, our Corollary 2 extends Theorem 1 by establishing SSC over an exponential time scale (as opposed to the diffusion scaling). For further perspective with respect to existing results, please refer to the discussion following the statement of Theorem 1 in Section 2.4.

5. Proof of Theorem 2

In this section, we present the proof of Theorem 2, organized in a sequence of four subsections. We first show in Section 5.1 that, for any network operating under a WMW policy, there is another network operating under an MW policy whose queue length process is a linear transformation of the queue length process of the original network. Thus, we can just focus on the MW policy. In Section 5.2, we review a general sensitivity result on a class of dynamical systems with piecewise constant drift. Next, in Section 5.3, we introduce an induced continuous-time dynamical system that provides the bridge between the original discrete-time process under an MW policy and the fluid model. The proof concludes in Section 5.4 by applying the general sensitivity result to the induced system.

5.1. From WMW to MW

In order to leverage the tools that we develop for MW policies and apply them to the more general WMW policies, we start with a reduction from WMW policies to an MW policy. This is accomplished through the following lemma, which shows that the queue lengths and fluid solutions under a WMW policy are linear transformations of queue lengths and fluid solutions under an MW policy in a transformed network.

Lemma 4 (Reduction of WMW Dynamics to MW Dynamics). *Consider a network N with action set $\mathcal{S}$ and a routing matrix R. Fix a weight vector w, an arrival function $A(\cdot)$, and an arrival rate vector λ. Let $Q(\cdot)$ be a queue length process of N corresponding to the arrival $A(\cdot)$ under a w-WMW policy. Let $W = \mathrm{diag}(w)$, $\widetilde{\lambda} = W^{1/2}\lambda$, and $\widetilde{A}(t) = W^{1/2}A(t)$ for all $t \in \mathbb{Z}_+$. Let $\widetilde{N}$ be a network with action set $\widetilde{\mathcal{S}} = W^{1/2}\mathcal{S}$ and routing matrix $\widetilde{R} = W^{1/2}RW^{-1/2}$. Then,*

a. *$\widetilde{Q}(t) = W^{1/2}Q(t)$ is a queue length process of $\widetilde{N}$ corresponding to the arrival $\widetilde{A}(\cdot)$ under an MW policy.*

b. *$\widetilde{q}(t) = W^{1/2}q(t)$ is a fluid solution of $\widetilde{N}$ corresponding to arrival rate $\widetilde{\lambda}$ and unit weights (as in MW) if and only if $q(t)$ is a fluid solution of N corresponding to arrival rate λ and WMW weights w.*

Proof. Given some $\mu \in \mathcal{S}$ and $Q \in \mathbb{R}_+^n$, we let $\widetilde{\mu} = W^{1/2}\mu$ and $\widetilde{Q} = W^{1/2}Q$. Then,

$$\begin{aligned}\widetilde{Q}^T\left(I - \widetilde{R}\right)\widetilde{\mu} &= \left(W^{1/2}Q\right)^T\left(I - W^{1/2}RW^{-1/2}\right)W^{1/2}\mu \\ &= Q^T W^{1/2}W^{1/2}(I - R)\,W^{-1/2}W^{1/2}\,\mu \\ &= Q^T W(I - R)\mu.\end{aligned}$$

Therefore, $\widetilde{\mu} \in \widetilde{\mathcal{S}}$ is a maximizer of $\widetilde{Q}(I - \widetilde{R})\widetilde{\mu}$ if and only if $\mu \in \mathcal{S}$ is a maximizer of $Q^T W(I - R)\mu$, that is, $\widetilde{\mathcal{S}}(\widetilde{Q}) = W^{1/2}\mathcal{S}_w(Q)$.

For part (a), for any $t \in \mathbb{Z}_+$,

$$\begin{aligned}\widetilde{Q}(t+1) &= W^{1/2}Q(t+1) \\ &= W^{1/2}Q(t) + W^{1/2}A(t) + W^{1/2}(R - I)\min\left(\mu(t), Q(t)\right) \\ &= \widetilde{Q}(t) + \widetilde{A}(t) + W^{1/2}(R - I)W^{-1/2}\,W^{1/2}\min\left(\mu(t), Q(t)\right) \\ &= \widetilde{Q}(t) + \widetilde{A}(t) + \left(\widetilde{R} - I\right)\min\left(\widetilde{\mu}(t), \widetilde{Q}(t)\right).\end{aligned}$$

Therefore, $\widetilde{Q}(\cdot)$ satisfies the evolution rule (3) of $\widetilde{N}$ and is a queue length process corresponding to the arrival function $\widetilde{A}(\cdot)$. Because $Q(t)$ evolves according to a w-WMW policy, we have $\mu(t) \in \mathcal{S}_w(Q(t))$. As shown earlier, this implies that $\widetilde{\mu}(t) \in \widetilde{\mathcal{S}}(\widetilde{Q})$, and thus, $\widetilde{Q}(t)$ indeed follows an MW policy.

For part (b), consider a set of functions $y(\cdot)$ and $s_\mu(\cdot)$ for $\mu \in \mathcal{S}$ that, together with $q(\cdot)$, satisfy (6)–(10). It is not difficult to see that all equations remain valid when $q(\cdot)$, λ, $\mathcal{S}$, $s_\mu(\cdot)$, $y(\cdot)$, and w are replaced with $\widetilde{q}(\cdot)$, $\widetilde{\lambda}$, $\widetilde{\mathcal{S}}$, $s_{\widetilde{\mu}}(\cdot) = s_\mu(\cdot)$, $W^{1/2}y(\cdot)$, and $\mathbb{1}_n$, respectively. The reverse direction is also true. Therefore, $\widetilde{q}(\cdot)$ is a fluid solution of $\widetilde{N}$ corresponding to the arrival rate vector $\widetilde{\lambda}$ with unit weights if and only if $q(\cdot)$ is a fluid solution of N corresponding to the arrival rate vector λ with weight vector w. □

5.2. Finitely Piecewise Constant Subgradient Dynamical Systems

In this section, we review some definitions and results from Sharifnassab et al. (2020). A dynamical system is identified with a set-valued function $F : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ and the associated differential inclusion $\dot{x}(t) \in F(x(t))$. We start with a formal definition, which allows for the presence of perturbations.

Definition 4 (Trajectories of a Dynamical System)**.** Consider a dynamical system $F : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ and let $U : \mathbb{R} \to \mathbb{R}^n$ be a right-continuous function, which we refer to as the *perturbation*. Suppose that $X(\cdot)$ and $\zeta(\cdot)$ are measurable functions of time that satisfy

$$X(t) = \int_0^t \zeta(\tau)\, d\tau + U(t), \qquad \forall\ t \geq 0, \tag{33}$$

$$\zeta(t) \in F(X(t)), \qquad \forall\ t \geq 0. \tag{34}$$

We then call X a *perturbed trajectory* corresponding to U. In the special case in which U is identically zero, we also refer to X as an *unperturbed trajectory*.

For a convex function $\Phi : \mathbb{R}^n \to \mathbb{R}$, we denote its subdifferential by $\partial\Phi(x)$. We say that F is a *subgradient dynamical system* if there exists a convex function $\Phi : \mathbb{R}^n \to \mathbb{R}$ such that, for any $x \in \mathbb{R}^n$, $F(x) = -\partial\Phi(x)$. Furthermore, if Φ is of the form

$$\Phi(x) = \max_i \left(-\mu_i^T x + b_i\right),$$

for some $\mu_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$, and with i ranging over a finite set, we say that F is a *finitely piecewise constant subgradient* (FPCS) system. Note that, for such systems, $F(x)$ is always equal to the convex hull of the vectors μ_i that maximize $-\mu_i^T x + b_i$.

FPCS systems admit a very special sensitivity bound.

Theorem 4 (Sharifnassab et al. (2020), theorem 1)**.** *Consider an FPCS system F. Then, there exists a constant C such that, for any unperturbed trajectory $x(\cdot)$ and for any perturbed trajectory $X(\cdot)$ with corresponding perturbation $U(\cdot)$ and the same initial conditions $X(0) = x(0)$, we have*

$$\left\|X(t) - x(t)\right\| \leq C \sup_{\tau \leq t} \left\|U(\tau)\right\|, \qquad \forall\ t \in \mathbb{R}_+. \tag{35}$$

Moreover, for any $\lambda \in \mathbb{R}^n$, the bound (35) applies to the (necessarily FPCS) system $F(\cdot) + \lambda$ with the same constant C.

5.3. Reduction of the MW Dynamics to an FPCS System

Throughout this section, we restrict attention to a network operated under an (unweighted) MW policy. In order to take advantage of Theorem 4, we show that a discrete-time network can also be represented as an associated (induced) FPCS dynamical system.

Definition 5 (Induced FPCS System)**.** For a network with action set $\mathcal{S}$ and routing matrix R, the induced FPCS system is the subgradient dynamical system F associated with the convex function

$$\Phi(x) = \max_{\mu \in \mathcal{S}} \left((I - R)\mu\right)^T x. \tag{36}$$

In particular, $F(x)$ is the convex hull of the image of $\mathcal{S}(x)$ under the linear transformation $R - I$, where $\mathcal{S}(x)$ is the set of vectors $\mu \in \mathcal{S}$ that maximize $((I - R)\mu)^T x$.

We start with the observation that fluid solutions of a network are trajectories of the induced FPCS system. Roughly speaking, this is because the service vectors chosen by the MW policy in (4) are maximizers of the set of linear functions $((I - R)\mu)^T Q$ over $\mu \in \mathcal{S}$, and the fluid solution moves along the negative of a convex combination of such maximizing service vectors. Thus, fluid solutions move along the subgradients of Φ [defined in (36)] and are, therefore, trajectories of the induced FPCS system.

Proposition 2 (Fluid Model Solutions as Trajectories of the Induced FPCS System)**.** *Consider a network and its induced FPCS system F. Let $q(\cdot)$ be a fluid solution of the network corresponding to arrival rate λ. Then, $q(\cdot)$ is an unperturbed trajectory of the dynamical system $\dot{q} \in F(q) + \lambda$. Conversely, any unperturbed trajectory $x(\cdot)$ of $F(\cdot) + \lambda$ with $x(0) \in \mathbb{R}_+^n$ is a fluid solution corresponding to λ.*

Proof. For a vector $\mu \in \mathbb{R}_+^n$ and a set $J \subseteq \{1, \ldots, n\}$ of indices, we let

$$D_J(\mu) \triangleq \left\{\xi \in \mathbb{R}_+^n \,\middle|\, \xi_i = \mu_i, \text{ for all } i \notin J, \text{and } 0 \leq \xi_j \leq \mu_j, \text{ for all } j \in J\right\}. \tag{37}$$

Equivalently,

$$D_J(\mu) \;=\; \mathrm{Conv}\big(\{\sigma_{-K}(\mu)\,\big|\,K\subseteq J\}\big), \tag{38}$$

where $\sigma_{-K}(\mu)$ is a vector whose ith entry is equal to the ith entry of μ if $i\notin K$ and equal to zero if $i\in K$. Recall that $\mathcal{S}(q)$ is defined as the set of all $\mu\in\mathcal{S}$ that maximize $((I-R)\mu)^T q$; see (4).

Claim 1. Fix a $q\in\mathbb{R}^n_+$ and a $\mu\in\mathcal{S}(q)$. Let $J=\{j\,|\,q_j=0\}$. Then,

$$(R-I)D_J(\mu)\subseteq F(q). \tag{39}$$

Proof of Claim 1. Note that, for any $\mu\in\mathcal{S}$ and any set K of indices, the vector $\sigma_{-K}(\mu)$ also belongs to $\mathcal{S}$ because of Assumption 1. We now fix some q and the set J as in the statement of the claim. For any $K\subseteq J$, we have $\sigma_{-K}(\mu)\preceq\mu$. Furthermore, because the entries of R and μ are nonnegative, we have $q^T R\,\sigma_{-K}(\mu)\le q^T R\,\mu$. Therefore,

$$\begin{aligned} q^T(I-R)\sigma_{-K}(\mu) \;&=\; q^T\sigma_{-K}(\mu)-q^T R\,\sigma_{-K}(\mu)\\ &=\; q^T\mu-q^T R\,\sigma_{-K}(\mu)\\ &\ge\; q^T\mu-q^T R\,\mu\\ &=\; q^T(I-R)\mu, \end{aligned}$$

where the second equality holds because $q_j=0$ whenever the jth entry of $\sigma_{-K}(\mu)$ is not equal to μ_j. We have, therefore, established that, if $\mu\in\mathcal{S}(q)$, then $\sigma_{-K}(\mu)\in\mathcal{S}(q)$. Because $F(q)$ is the image under $R-I$ of the convex hull of $\mathcal{S}(q)$, we obtain

$$(R-I)\sigma_{-K}(\mu)\in F(q). \tag{40}$$

Therefore,

$$\begin{aligned} (R-I)D_J(\mu) &= (R-I)\,\mathrm{Conv}\big(\{\sigma_{-K}(\mu)\,\big|\,K\subseteq J\}\big)\\ &= \mathrm{Conv}\big(\{(R-I)\sigma_{-K}(\mu)\,\big|\,K\subseteq J\}\big)\\ &\subseteq\; F(q), \end{aligned}$$

where the first equality is due to (38), and the last relation is due to (40) and the convexity of $F(q)$. This establishes the validity of the claim (39). □

We now return to the proof of the proposition. Consider a function $y(\cdot)$ and a set of functions $s_\mu(\cdot)$, $\mu\in\mathcal{S}$, that, together with $q(\cdot)$, satisfy the fluid model Equations (6)–(10). We show that $q(\cdot)$ satisfies the differential inclusion $\dot q\in F(q)+\lambda$. Fix some $t\ge 0$ and let $y=y(t)$. For any $\mu\in\mathcal{S}$, let $s_\mu=s_\mu(t)$ and consider an n-dimensional vector y^μ with entries

$$y_i^\mu=\begin{cases} y_i\mu_i/\sum_{\nu\in\mathcal{S}} s_\nu\nu_i, & \text{if } \sum_{\nu\in\mathcal{S}} s_\nu\nu_i\neq 0,\\ 0, & \text{otherwise.}\end{cases}$$

It follows that, for $i=1,\dots,n$,

$$\sum_{\mu\in\mathcal{S}} s_\mu y_i^\mu=\sum_{\mu\in\mathcal{S}} s_\mu\frac{y_i\mu_i}{\sum_{\nu\in\mathcal{S}}s_\nu\nu_i}\;=\;y_i. \tag{41}$$

Then,

$$\sum_{\mu\in\mathcal{S}} s_\mu y^\mu=y. \tag{42}$$

On the other hand, for any $\mu\in\mathcal{S}$ and for $i=1,\dots,n$, either $y_i^\mu=0$ or (8) implies that $y_i^\mu=\mu_i(y_i/\sum_{\nu\in\mathcal{S}}s_\nu\nu_i)\le\mu_i$. Moreover, for any $\mu\in\mathcal{S}$ and any $i\le n$, if $q_i(t)>0$, then, from (9), $y_i^\mu=y_i(\mu_i/\sum_{\nu\in\mathcal{S}}s_\nu\nu_i)=0$. Letting $J=\{j\,|\,q_j(t)=0\}$, it then follows from the definition of $D_J(\mu)$ that, for any $\mu\in\mathcal{S}$, $\mu-y^\mu\in D_J(\mu)$. Claim 1 then implies that $(R-I)(\mu-y^\mu)\in F(q(t))$. Therefore, in light of (7) and the convexity of $F(q(t))$, we have

$$\sum_{\mu\in\mathcal{S}} s_\mu(R-I)\big(\mu-y^\mu\big)\;\in\;F\big(q(t)\big). \tag{43}$$

Finally, from (6),

$$\begin{aligned}\dot{q}(t) &= \lambda + (R-I)\left(\sum_{\mu\in\mathcal{S}} s_\mu \mu - y\right)\\ &= \lambda + (R-I)\left(\sum_{\mu\in\mathcal{S}} s_\mu \mu - \sum_{\mu\in\mathcal{S}} s_\mu y^\mu\right)\\ &= \lambda + \sum_{\mu\in\mathcal{S}} s_\mu (R-I)\big(\mu - y^\mu\big)\\ &\in F\big(q(t)\big) + \lambda,\end{aligned} \tag{44}$$

where the second equality is due to (42).

We now prove the converse part of the proposition, that every unperturbed trajectory $x(\cdot)$ of $F(\cdot)+\lambda$ initialized in the positive orthant is a fluid solution. Consider a fluid solution $q(\cdot)$ corresponding to the arrival rate λ, initialized with $q(0) = x(0)$ (for proofs of existence, see appendix A of Dai and Lin (2005) and lemma 9 of Markakis et al. (2018)).

It then follows from the first part of this proposition that $q(\cdot)$ is also an unperturbed trajectory of $F(\cdot)+\lambda$. On the other hand, it is shown in Rockafellar (1970) that any subgradient dynamical system is a maximal monotone map.[13] Then, corollary 4.6 of Stewart (2011) implies that there is a unique unperturbed trajectory with initial point $x(0)$. Therefore, $x(t) = q(t)$ for all $t \geq 0$, and the desired result follows. □

Proposition 2 has established that a fluid solution is a trajectory of the induced FPCS system. We now show, in the next proposition, that the actual discrete-time queue length process is close to a perturbed trajectory of the induced FPCS system. Note that, even if the discrete-time system has completely deterministic and steady arrivals (no stochastic fluctuations) it can still "chatter" around the boundary separating two regions with different drifts. The idea behind the proof is that this chattering can also be viewed as a perturbation of a straight trajectory. This is conceptually straightforward, but some of the details of the behavior in the vicinity of such boundaries are tedious.

Proposition 3 (Queue Length Processes as Trajectories of the Induced FPCS System)**.** *For any network, there exists a constant β that satisfies the following statement. Fix a $\lambda \in \mathbb{R}_+^n$ and let $A(\cdot)$ be an arrival function and $Q(\cdot)$ be a corresponding queue length process. Then, there exists a right-continuous (perturbation) function $U(\cdot) : \mathbb{R}_+ \to \mathbb{R}^n$ satisfying*

$$\sup_{\tau\leq t}\big\|U(\tau)\big\| \leq \big\|\lambda\big\| + \beta + \sup_{\tau\leq t}\left\|\sum_{\substack{k<\tau\\ k\in\mathbb{Z}_+}} (A(k)-\lambda)\right\|, \qquad \forall\ t\in\mathbb{R}_+, \tag{45}$$

and a corresponding perturbed trajectory $X(\cdot)$ of $F(\cdot)+\lambda$ such that

$$\big\|X(k) - Q(k)\big\| \leq \beta, \quad \forall k \in \mathbb{Z}_+. \tag{46}$$

It is possible to strengthen Proposition 3 and ensure that we actually have $X(k) = Q(k)$ for every $k \in \mathbb{Z}_+$, thus strengthening (46). However, this stronger result is not needed for our future development and would require a much more tedious construction of $U(\cdot)$. It is also worth pointing out here that the perturbed trajectory $X(\cdot)$ in our construction is always nonnegative.

Before presenting the detailed proof, we provide some intuition on the issues that arise. Recall the network evolution rule $Q(k+1) = Q(k) + A(k) + (R-I)\min(\mu(k), Q(k))$ in (3). Consider a time k at which there is a unique maximizer $\mu(k)$ so that $F(Q(k))$ is a singleton, consisting of the single element $(R-I)\mu(k)$. Suppose, furthermore, that $\mu(k) \preceq Q(k)$. In this case, (3) becomes $Q(k+1) = Q(k) + A(k) + (R-I)\mu(k)$. Let

$$U(t) = \begin{cases} -(t-k)\big[(R-I)\mu(k) + \lambda\big], & \text{for} \quad t\in(k,k+1),\\ A(k) - \lambda, & \text{for} \quad t = k+1.\end{cases}$$

In the interval $t \in (k, k+1)$, we have $\dot{U}(t) = -[(R-I)\mu(k) + \lambda] \in -F(Q(k)) - \lambda$. Suppose now that $X(k) = Q(k)$. From the dynamics of $X(\cdot)$ (see Definition 4), we have

$$\dot{X}(t) = F(X(t)) + \lambda + \dot{U}(t) = F(Q(k)) + \lambda + \dot{U}(t) = 0,$$

and the perturbed trajectory remains constant: $X(t) = Q(k)$ for $t \in (k, k+1)$. Finally, a discontinuity in $U(\cdot)$ at $t = k+1$ forces $X(t)$ to jump to the new value $Q(k+1)$. Thus, in this example, we have a perturbed trajectory that agrees with the queue process at integer times.

This argument, however, fails when $\min(\mu(k), Q(k)) \neq \mu(k)$ because the received service in time slot k, that is, $(R-I)\min(\mu(k), Q(k))$, needs not belong to $F(Q(k))$. To circumvent this problem, we find a nearby point $y(k)$ such that $(R-I)\min(\mu(k), Q(k)) \in F(y(k))$. We then construct $U(\cdot)$ so that it forces $X(t)$ to jump to $y(k)$ at time $t = k$ and stay there for $t \in [k, k+1)$.

Proof. We now provide the detailed proof of Proposition 3. We make use of the following known result:

Lemma 5 (Mannor and Tsitsiklis (2005), lemma 5.1)**.** *Given a finite collection of half spaces $H_i \subset \mathbb{R}^n$ with nonempty intersection, there exists a constant $c > 0$ such that*

$$d\left(x, \bigcap_i H_i\right) \leq c \cdot \max_i d(x, H_i), \qquad \forall\, x \in \mathbb{R}^n. \tag{47}$$

We begin with a claim.

Claim 2. There exists a constant κ that satisfies the following. Consider any $x \in \mathbb{R}^n_+$ and any $\mu \in \mathcal{S}(x)$. Let $J = \{j \,|\, x_j \leq \mu_j\}$. Then, there exists a $y \in \mathbb{R}^n_+$ such that $\|y - x\| \leq \kappa$ and

$$(R-I)D_J(\mu) \subseteq F(y), \tag{48}$$

where $D_J(\mu)$ is defined in (37).

Proof of Claim 2. We leverage Lemma 5 to find y and then use Claim 1 to prove (48). To every $\mu \in \mathcal{S}$, we associate an effective region R_μ:

$$R_\mu = \{z \,|\, \mu \in \mathcal{S}(z)\}. \tag{49}$$

Fix some $x \in \mathbb{R}^n_+$ and some $\mu \in \mathcal{S}(x)$. The effective region R_μ is then the intersection of half spaces of the form

$$H_\pi = \left\{z \in \mathbb{R}^n \,\middle|\, z^T(I-R)\mu \geq z^T(I-R)\pi\right\}, \qquad \pi \in \mathcal{S}. \tag{50}$$

Because $\mu \in \mathcal{S}(x)$, we have $x \in H_\pi$ and $d(x, H_\pi) = 0$ for all $\pi \in \mathcal{S}$. Let

$$b \triangleq \max_{\pi \in \mathcal{S}} \max_{i \leq n} \pi_i \tag{51}$$

be the maximum service capacity of any queue over all service vectors.

We also define, for every $i \leq n$, two half spaces $H_{j^+} = \{z \in \mathbb{R}^n \,|\, z_j \geq 0\}$ and $H_{j^-} = \{z \in \mathbb{R}^n \,|\, z_j \leq 0\}$. It follows from the definition of J that, for any $j \in J$, $d(x, H_{j^-}) \leq b$, and from $x \in \mathbb{R}^n_+$ that $d(x, H_{i^+}) = 0$ for all $i \leq n$. We define a set B, which is determined by the chosen $x \in \mathbb{R}^n_+$ and $\mu \in \mathcal{S}(x)$, as follows:

$$B = \{z \in \mathbb{R}^n_+ \cap R_\mu \,|\, z_j = 0 \text{ whenever } x_j \leq \mu_j\}.$$

Note that the set B is the intersection of finitely many half spaces of the form H_π, H_{j^+}, and H_{j^-}. Note, furthermore, that B contains the origin and is, therefore, nonempty. Finally note that x has a distance of at most b from each of the half spaces defining B. Therefore, Lemma 5 implies that

$$d(x, B) \leq c\, b, \tag{52}$$

for some constant c. In general, the constant c depends on the particular x and μ under consideration. Note, however, that the set B is completely determined by $\mu \in \mathcal{S}$ and the set of indices $J = \{j \,|\, x_j \leq \mu_j\}$. There are finitely many choices for μ and for J, hence, finitely many possible sets B. By taking the largest of the constants c associated with different sets B, we see that c in (52) can be taken to be an absolute constant, independent of x and μ.

Let y be the closest point to x in the set B. Letting $\kappa = cb$, (52) implies that

$$\|y - x\| \leq \kappa, \tag{53}$$

where κ is an absolute constant.

Because $y \in B$, we have $y \in R_\mu$ so that $\mu \in S(y)$. Moreover, for any $j \in J$, that is, if $x_j \le \mu_j$, we have $y_j = 0$. Let $J' = \{j \mid y_j = 0\}$. We then have $J \subseteq J'$. We now apply this inclusion together with Claim 1 with y and J' playing the role of q and J in the statement of the claim to obtain

$$(R - I)D_J(\mu) \subseteq (R - I)D_{J'}(\mu) \subseteq F(y).$$

This completes the proof of the claim. □

For every $t \in \mathbb{Z}_+$, let $\mu(t) \in \mathcal{S}(Q(t))$ be the action taken by the scheduler at time t, $J(t) = \{j \mid Q_j(t) \le \mu_j(t)\}$, and $y(t) \in \mathbb{R}_+^n$ be a vector that satisfies $\|y(t) - Q(t)\| \le \kappa$ and

$$(R - I)D_J(\mu(t)) \subseteq F(y(t)), \tag{54}$$

as in Claim 2.

We now proceed to the main part of the proof of the proposition. With a slight abuse of notation, we write expressions such as $\sum_{k<t}$ even if t is noninteger, which we interpret as $\sum_{\{k \in \mathbb{Z}_+ \mid k<t\}}$. We define the right-continuous perturbation function $U(\cdot)$ as

$$\begin{aligned} U(t) = &\sum_{k \le t-1} (A(k) - \lambda) + y(\lfloor t \rfloor) - Q(\lfloor t \rfloor) \\ &- (t - \lfloor t \rfloor)\big(\lambda + (R - I)\min\big(\mu(\lfloor t \rfloor), Q(\lfloor t \rfloor)\big)\big), \end{aligned} \tag{55}$$

for all $t \in \mathbb{R}_+$.

Let

$$b \triangleq \max_{\mu \in \mathcal{S}} \sup_{Q \in \mathbb{R}_+^n} \big\|(R - I)\min(\mu, Q)\big\|, \tag{56}$$

which is a finite constant. For any $t \in \mathbb{R}_+$,

$$\begin{aligned} \sup_{\tau \le t} \|U(\tau)\| &\le \sup_{\tau \le t} \left\| \sum_{k<\tau} (A(k) - \lambda) \right\| + \sup_{k \le \tau} \|y(k) - Q(k)\| + \|\lambda\| \\ &\quad + \sup_{k \le \tau} \big\|(R - I)\min\big(\mu(k), Q(k)\big)\big\| \\ &\le \sup_{\tau \le t} \left\| \sum_{k<\tau} (A(k) - \lambda) \right\| + \kappa + \|\lambda\| + b. \end{aligned} \tag{57}$$

Hence, (45) follows for $\beta = \kappa + b$.

For every $t \in \mathbb{R}_+$, let

$$X(t) = y(\lfloor t \rfloor). \tag{58}$$

Because $\|y(t) - Q(t)\| \le \kappa$, by construction, the desired equality (46) at integer times is trivially true with κ playing the role of β. It remains to show that $X(\cdot)$ is a perturbed trajectory.

Let

$$\xi(t) = (R - I)\min\big(\mu(\lfloor t \rfloor), Q(\lfloor t \rfloor)\big) + \lambda, \qquad \forall\ t \in \mathbb{R}_+. \tag{59}$$

Because $\min(\mu(\lfloor t \rfloor), Q(\lfloor t \rfloor)) \in D_{J(\lfloor t \rfloor)}(\mu(\lfloor t \rfloor))$, it follows from (54) and (58) that $\xi(t) \in F(X(t)) + \lambda$ for all $t \in \mathbb{R}_+$.

We now show that

$$X(t) = \int_0^t \xi(\tau)\, d\tau + U(t), \qquad \forall t \ge 0. \tag{60}$$

For any $k \in \mathbb{Z}_+$,

$$\begin{aligned}\int_k^{k+1} \xi(\tau)\,d\tau + U(k+1) - U(k) &= \big[(R-I)\min\big(\mu(k), Q(k)\big) + \lambda\big] \\ &\quad + \big[A(k) - \lambda + y(k+1) - Q(k+1) \\ &\quad\quad - \big(y(k) - Q(k)\big)\big] \\ &= \big[(R-I)\min\big(\mu(k), Q(k)\big) + A(k)\big] \\ &\quad - \big(Q(k+1) - Q(k)\big) + \big(y(k+1) - y(k)\big) \\ &= y(k+1) - y(k) \\ &= X(k+1) - X(k),\end{aligned} \tag{61}$$

where the third equality follows from the evolution rule (3). Moreover, for any $t \notin \mathbb{Z}$,

$$\begin{aligned}\int_{\lfloor t\rfloor}^{t} \xi(\tau)\,d\tau + U(t) - U(\lfloor t\rfloor) &= (t - \lfloor t\rfloor)\big(\lambda + (R-I)\min\big(\mu(\lfloor t\rfloor), Q(\lfloor t\rfloor)\big)\big) \\ &\quad - (t - \lfloor t\rfloor)\big(\lambda + (R-I)\min\big(\mu(\lfloor t\rfloor), Q(\lfloor t\rfloor)\big)\big) \\ &= 0 \\ &= X(t) - X(\lfloor t\rfloor).\end{aligned} \tag{62}$$

Then, a simple induction based on (61) and (62) implies (60), and therefore, $X(\cdot)$ is a perturbed trajectory of $F(\cdot) + \lambda$ corresponding to $U(\cdot)$, which is the desired result. □

5.4. Proof of Theorem 2

Having established a reduction from WMW policies to an MW policy (in Lemma 4) and a reduction from a network, operating under MW policy, to its induced FPCS system (see Propositions 2 and 3), we can now leverage Theorem 4 to prove Theorem 2.

Proof of Theorem 2. Consider a network N that operates under a w-WMW policy. Let $W = \mathrm{diag}(w)$. Consider a queue length process $Q(\cdot)$ of N corresponding to an arrival $A(\cdot)$ and a fluid solution $q(\cdot)$ of N corresponding to an arrival rate vector λ initialized at $q(0) = Q(0)$. For each time t, let $\widetilde{Q}(t) = W^{1/2}Q(t)$, $\widetilde{q}(t) = W^{1/2}q(t)$, $\widetilde{A}(t) = W^{1/2}A(t)$, and $\widetilde{\lambda} = W^{1/2}\lambda$. Let $\theta_{\min} \triangleq \min_i w_i^{1/2}$ and $\theta_{\max} \triangleq \max_i w_i^{1/2}$. Then, for any time $t \in \mathbb{Z}_+$,

$$\big\|\widetilde{Q}(t) - \widetilde{q}(t)\big\| = \big\|W^{1/2}\big(Q(t) - q(t)\big)\big\| \;\geq\; \theta_{\min}\big\|Q(t) - q(t)\big\|, \tag{63}$$

and

$$\left\|\sum_{k\leq t}\Big(\widetilde{A}(k) - \widetilde{\lambda}\Big)\right\| = \left\|W^{1/2}\sum_{k\leq t}\big(A(k) - \lambda\big)\right\| \;\leq\; \theta_{\max}\left\|\sum_{k\leq t}\big(A(k) - \lambda\big)\right\|. \tag{64}$$

Let $\widetilde{N}$ be, as in Lemma 4, a network that operates under an MW policy and for which $\widetilde{Q}(\cdot)$ and $\widetilde{q}(\cdot)$ are a queue length process and a fluid solution, respectively. Consider the induced FPCS system F of $\widetilde{N}$. It follows from Proposition 3 that there exists a right-continuous perturbation function $U(\cdot)$ satisfying, for any $t \in \mathbb{R}_+$,

$$\sup_{\tau\leq t}\big\|U(\tau)\big\| \;\leq\; \big\|\widetilde{\lambda}\big\| + \beta + \sup_{\tau\leq t}\left\|\sum_{k<\tau}\Big(\widetilde{A}(k) - \widetilde{\lambda}\Big)\right\|, \tag{65}$$

and a corresponding perturbed trajectory $X(\cdot)$ of $F(\cdot) + \widetilde{\lambda}$ such that

$$\big\|X(k) - \widetilde{Q}(k)\big\| \leq \beta, \quad \forall k \in \mathbb{Z}_+, \tag{66}$$

where β is a constant independent of $\widetilde{\lambda}$. Moreover, from Proposition 2, $\widetilde{q}(\cdot)$ is an unperturbed trajectory of $F(\cdot) + \widetilde{\lambda}$. Then, applying Theorem 4 for the FPCS sytem F, we obtain, for any $t \in \mathbb{R}_+$,

$$\big\|X(t) - \widetilde{q}(t)\big\| \;\leq\; \widetilde{C}\sup_{\tau\leq t}\big\|U(\tau)\big\|, \tag{67}$$

for some constant $\widetilde{C} \geq 1$ that is independent of λ. Let $C = \widetilde{C}\,\max(\theta_{\max}, 2\beta)\,/\theta_{\min}$. Then, for any $t \in \mathbb{Z}_+$,

$$\begin{aligned}
\|Q(t) - q(t)\| &\leq \frac{1}{\theta_{\min}} \|\widetilde{Q}(t) - \widetilde{q}(t)\| \\
&\leq \frac{1}{\theta_{\min}} \left(\|X(t) - \widetilde{q}(t)\| + \beta \right) \\
&\leq \frac{1}{\theta_{\min}} \left(\widetilde{C} \sup_{\tau \leq t} \|U(\tau)\| + \beta \right) \\
&\leq \frac{\widetilde{C}}{\theta_{\min}} \left(\sup_{\tau \leq t} \|U(\tau)\| + \beta \right) \\
&\leq \frac{\widetilde{C}}{\theta_{\min}} \left(\|\widetilde{\lambda}\| + 2\beta + \sup_{\tau \leq t} \Big\| \sum_{k<\tau} \big(\widetilde{A}(k) - \widetilde{\lambda} \big) \Big\| \right) \\
&\leq \frac{\widetilde{C}}{\theta_{\min}} \left(\theta_{\max} \|\lambda\| + 2\beta + \theta_{\max} \sup_{\tau \leq t} \Big\| \sum_{k<\tau} (A(k) - \lambda) \Big\| \right) \\
&\leq C \left(1 + \|\lambda\| + \sup_{\tau \leq t} \Big\| \sum_{k<\tau} (A(k) - \lambda) \Big\| \right),
\end{aligned}$$

where the relations are due to (63), (66), (67), $\widetilde{C} \geq 1$, (65), (64), and the definition of C, respectively. □

6. Proof of Theorem 3

In this section, we present the proof of Theorem 3. Part (a) is a corollary of part (b). In the following, we first prove part (b) and then part (c).

Proof of Part (b). We first consider an MW policy. We then use Lemma 4 to extend the result to the case of WMW policies. The proof for an MW policy goes along the following lines. We break down a $g(r)$-long interval into subintervals of length r. We define a "good" event $\mathscr{E}_r$ such that the aggregate arrival in each r-long interval does not deviate much from its average and show that this event happens with high probability. We then use Theorem 2 to show that $\mathscr{E}_r$ implies that the queue length process stays close to a fluid solution in every subinterval. These fluid solutions are attracted to $\mathcal{I}(\lambda)$ (see Lemma 2) and, hence, also keep the queue length process near $\mathcal{I}(\lambda)$.

We now present a detailed proof. We fix some $\delta > 0$ and some $\epsilon > 0$ such that

$$\epsilon \leq \min(\delta, \alpha)/4C, \tag{68}$$

where C is the sensitivity constant provided by Theorem 2 and $\alpha = \alpha(\lambda)$ is the constant in Lemma 2 associated with λ. For $r \in \mathbb{N}$, we define a good event $\mathscr{E}_r$ as follows:

$$\mathscr{E}_r = \left\{ \frac{1}{r} \cdot \sup_{t \in [ir, (i+1)r)} \Big\| \sum_{\tau=ir}^{t} (A^r(\tau) - \lambda^r) \Big\| \leq \epsilon, \quad \forall\, i \in \big[0,\, g(r)T/r\big] \right\}. \tag{69}$$

We denote the complement of an event $\mathscr{E}$ by $\mathscr{E}^c$. Then, for any r,

$$\begin{aligned}
\mathrm{P}\big(\mathscr{E}_r^c\big) &= \mathrm{P}\left(\exists\, i \in [0,\, g(r)T/r],\ \text{s.t.}\ \frac{1}{r} \sup_{t \in [ir, (i+1)r)} \Big\| \sum_{\tau=ir}^{t} (A^r(\tau) - \lambda^r) \Big\| > \epsilon \right) \\
&\leq \sum_{i=0}^{\lfloor g(r)T/r \rfloor} \mathrm{P}\left(\frac{1}{r} \sup_{t \in [ir, (i+1)r)} \Big\| \sum_{\tau=ir}^{t} (A^r(\tau) - \lambda^r) \Big\| > \epsilon \right) \\
&= \big(\lfloor g(r)T/r \rfloor + 1 \big)\, \mathrm{P}\left(\frac{1}{r} \sup_{t \in [0, r)} \Big\| \sum_{\tau=0}^{t} (A^r(\tau) - \lambda^r) \Big\| > \epsilon \right) \\
&= \big(\lfloor g(r)T/r \rfloor + 1 \big)\, o\big(1/f(r, \epsilon)\big),
\end{aligned} \tag{70}$$

where the inequality is due to the union bound, the second equality holds because $A^r(\cdot)$ is a stationary process, and the last equality is because $A^r(\cdot)$, $r \in \mathbb{N}$, is an f-tailed sequence of processes (see Definition 2). Also note that, in the last line, ϵ is a fixed constant, and the $o(\cdot)$ notation is with respect to r as r goes to infinity.

From now on and because λ is fixed, we use the simpler notation $\mathcal{J}$ instead of $\mathcal{J}(\lambda)$. Consider an $r_0 \in \mathbb{N}$ such that, for every $r \geq r_0$,

$$\|\lambda^r - \lambda\| \leq C\epsilon \tag{71}$$

$$d\big(\widehat{q}^r(0)\,,\,\mathcal{J}\big) \leq 2C\epsilon, \tag{72}$$

$$\lambda^r + 1 \leq r\epsilon. \tag{73}$$

Such an r_0 exists because of the convergence assumptions in the statement of the theorem, which also imply that λ^r is a bounded sequence. For every $r, i \in \mathbb{Z}_+$, we define the following two events, $E_{r,i}$ and $E'_{r,i}$:

$$\begin{aligned} E_{r,i} &\triangleq \text{the event that } \; d(Q^r(ir)\,,\,\mathcal{J}) \leq 2Cr\epsilon, \\ E'_{r,i} &\triangleq \text{the event that } \; d(Q^r(t)\,,\,\mathcal{J}) \leq r\delta, \quad \forall\, t \in [ir, (i+1)r), \end{aligned} \tag{74}$$

where $Q^r(\cdot)$ is the queue length process corresponding to the arrival $A^r(\cdot)$. Using Theorem 2, we now show that, for any $r \geq r_0$, $\mathscr{E}_r$ implies $E_{r,i}$ and $E_{r,i}$ for all $i < g(r)T/r$.

Claim 3. Fix some $r \geq r_0$. The occurrence of the event $\mathscr{E}_r$ implies the occurrence of the events $E_{r,i}$ and $E'_{r,i}$ for all $i < g(r)T/r$.

Proof of Claim 3. The proof is by induction on i. For the base case, $E_{r,0}$ follows from (72), because of $Q^r(0) = r\widehat{q}^r(0)$ and the conic property of $\mathcal{J}$ in (13). For the induction step, we show that, for any $i < g(r)T/r$, the events $\mathscr{E}_r$ and $E_{r,i}$ imply $E_{r,i+1}$ and $E'_{r,i}$.

For any $r \in \mathbb{Z}_+$, let $q_i^r(t)$ be the fluid solution corresponding to arrival rate λ^r and initialized with $q_i^r(ir) = Q^r(ir)$ at time ir. Fix an arbitrary $t_0 \geq ir$ and let $q(\cdot)$ be a fluid solution corresponding to arrival rate λ, initialized at $q(t_0) = q_i^r(t_0)$. From Proposition 2, $q(\cdot)$ and $q_i^r(\cdot)$ are solutions of $\dot{q} \in F(q) + \lambda$ and $\dot{q}_i^r \in F(q_i^r) + \lambda^r$, respectively, where F is the induced FPCS system of the network. It then follows from lemma 4.5 of Stewart (2011) that, for any $t \geq t_0$, $\|q_i^r(t) - q(t)\| \leq (t - t_0)\|\lambda^r - \lambda\|$. As a result,

$$\frac{d^+}{dt}\big\|q_i^r(t) - q(t)\big\|\,\Big|_{t=t_0} \;\leq\; \|\lambda^r - \lambda\|. \tag{75}$$

Suppose that $q_i^r(t_0) \notin \mathcal{J}$. Then, we also have $q(t_0) \notin \mathcal{J}$, and Lemma 2 implies that

$$\begin{aligned} \frac{d^+}{dt} d\big(q_i^r(t), \mathcal{J}\big)\Big|_{t_0} \;&\leq\; \frac{d^+}{dt}\big\|q_i^r(t) - q(t)\big\|\Big|_{t=t_0} + \frac{d^+}{dt} d\big(q(t), \mathcal{J}\big)\Big|_{t=t_0} \\ &\leq\; \|\lambda^r - \lambda\| - \alpha \\ &\leq\; C\epsilon - \alpha \\ &\leq\; C\epsilon - 4C\epsilon \\ &<\; -2C\epsilon, \end{aligned} \tag{76}$$

where the second inequality is from (75), and the fourth inequality is due to (68). Moreover, under $E_{r,i}$, we have

$$d\big(q_i^r(ir),\, \mathcal{J}\big) \;= d(Q^r(ir),\, \mathcal{J}) \leq 2Cr\epsilon.$$

Therefore, under $E_{r,i}$,

$$d\big(q_i^r(t),\, \mathcal{J}\big) \;\leq\; 2Cr\epsilon, \qquad \forall\, t \in [ir,\, (i+1)r), \tag{77}$$

$$d\big(q_i^r((i+1)r),\, \mathcal{J}\big) = 0. \tag{78}$$

Then, for any $r, i \in \mathbb{Z}_+$ and under $E_{r,i}$,

$$\begin{aligned} d(Q^r((i+1)r),\, \mathcal{J}) \;&\leq\; \big\|Q^r((i+1)r) - q_i^r((i+1)r)\big\| + d\big(q_i^r((i+1)r),\, \mathcal{J}\big) \\ &=\; \big\|Q^r((i+1)r) - q_i^r((i+1)r)\big\| \\ &\leq\; C\left(1 + \lambda^r + \sup_{t \in [ir, (i+1)r)} \left\|\sum_{\tau=ir}^{t} (A^r(\tau) - \lambda^r)\right\|\right) \\ &\leq\; C(r\epsilon + r\epsilon) \\ &=\; 2Cr\epsilon, \end{aligned}$$

where the second inequality is due to Theorem 2 and the last inequality follows from (73) and $\mathscr{E}_r$. This implies $E_{r,i+1}$. Moreover,

$$\begin{aligned}\sup_{t\in[ir,(i+1)r)} d(Q^r(t),\mathcal{F}) &\le \sup_{t\in[ir,(i+1)r)} \left(\left\|Q^r(t)-q_i^r(t)\right\| + d\left(q_i^r(t),\mathcal{F}\right)\right)\\ &\le \sup_{t\in[ir,(i+1)r)} \left\|Q^r(t)-q_i^r(t)\right\| + 2Cr\epsilon\\ &\le C\left(1+\lambda^r + \sup_{t\in[ir,(i+1)r)} \left\|\sum_{\tau=0}^{t-1}(A^r(\tau)-\lambda^r)\right\|\right) + 2Cr\epsilon\\ &\le C(r\epsilon + r\epsilon) + 2Cr\epsilon\\ &\le r\delta,\end{aligned}$$

where the second inequality is due to (77), the third inequality follows from Theorem 2, the fourth inequality is from (73) and $\mathscr{E}_r$, and the last inequality is due to the definition of ϵ in (68). This implies $E'_{r,i}$ and completes the proof of the claim. □

Back to the proof of the theorem, let us again fix some $r \ge r_0$. We have

$$\begin{aligned}\mathrm{P}\left(\sup_{t\in[0,T]} d\left(\widehat{q}^r(t)\,,\,\mathcal{F}\right) > \delta\right) &= \mathrm{P}\left(\sup_{t\le g(r)T} d(Q^r(t)/r\,,\,\mathcal{F}) > \delta\right)\\ &= \mathrm{P}\left(\sup_{t\le g(r)T} d(Q^r(t)\,,\,\mathcal{F}) > r\delta\right)\\ &\le \mathrm{P}\left(\bigcup_{i\le g(r)T/r} {E'}^c_{r,i}\right)\\ &\le \mathrm{P}\left(\mathscr{E}_r^c\right),\end{aligned} \tag{79}$$

where the second equality is due to (13), the first inequality is from the definition of $E'_{r,i}$, and the last inequality is due to Claim 3. Thus,

$$\begin{aligned}\frac{rf(r,\epsilon)}{g(r)}\,\mathrm{P}\left(\sup_{t\in[0,T]} d\left(\widehat{q}^r(t)\,,\,\mathcal{F}\right) > \delta\right) &\le \frac{rf(r,\epsilon)}{g(r)}\,\mathrm{P}(\mathscr{E}^c)\\ &\le \frac{rf(r,\epsilon)}{g(r)}\left(\lfloor g(r)T/r\rfloor + 1\right) o\left(1/f(r,\epsilon)\right)\\ &\le o\left(\frac{\lfloor g(r)T/r\rfloor + 1}{g(r)/r}\right)\\ &= o(1) \underset{r\to\infty}{\longrightarrow} 0,\end{aligned}$$

where the first two inequalities are due to (79) and (70), respectively, and the equality follows from the assumption $\liminf_{r\to\infty} g(r)/r > 0$. This completes the proof of part (b) for the case of an MW policy.

We now present the proof of part (b) for WMW policies. Suppose that a network N operates under a w-WMW policy and consider an associated network $\widetilde{N}$ as in Lemma 4 along with the variables and processes therein. It follows from Lemma 4 that, if the constant function $q(t) = q_0$ is a fluid solution for network N, then the constant function $\widetilde{q}(t) = W^{1/2}q_0$ is a fluid solution for $\widetilde{N}$ under an MW policy and vice versa. Therefore, $\widetilde{\mathcal{F}} = W^{1/2}\mathcal{F}$ is the set of invariant states for $\widetilde{N}$, corresponding to arrival rate $\widetilde{\lambda} = W^{1/2}\lambda$. Let $\theta_{\min} \triangleq \min_i w_i^{1/2}$ and $\theta_{\max} \triangleq \max_i {w_i}^{1/2}$. Let $\widetilde{q}(\cdot) = W^{1/2}\widehat{q}(\cdot)$, which is the scaled version of the MW-driven process $\widetilde{Q}(\cdot)$. Then, for any $r \in \mathbb{N}$ and any time t,

$$d\left(\widehat{q}^r(t)\,,\,\mathcal{F}\right) = d\left(W^{-1/2}\widetilde{q}^r(t)\,,\,W^{-1/2}\widetilde{\mathcal{F}}\right) \le \frac{1}{\theta_{\min}} d\left(\widetilde{q}^r(t)\,,\,\widetilde{\mathcal{F}}\right). \tag{80}$$

In the same vein,

$$\left\|\widetilde{A}^r(t) - \widetilde{\lambda}^r\right\| = \left\|W^{1/2}(A^r(t)-\lambda^r)\right\| \le \theta_{\max}\left\|A^r(t)-\lambda^r\right\|. \tag{81}$$

As a result, $\widetilde{A}^r(\cdot)$, $r \in \mathbb{N}$, is a $(\theta_{\max} f)$-tailed sequence of processes.

As in (68), fix some $\delta > 0$ and some $\epsilon > 0$ such that $\epsilon \le \min(\delta\theta_{\min}, \alpha)/4C$, where C is the sensitivity constant of the network operating under a MW policy (see Theorem 2) and $\alpha = \alpha(\lambda)$ is the constant in Lemma 2 associated with λ. Using what we have already established for MW policies, it follows that

$$\frac{r\theta_{\max} f(r,\epsilon)}{g(r)} \mathrm{P}\left(\sup_{t\in[0,T]} d\left(\widetilde{q}^r(t), \widetilde{\mathcal{F}}\right) > \delta\theta_{\min} \right) \underset{r\to\infty}{\longrightarrow} 0. \tag{82}$$

This, together with (80), implies that

$$\begin{aligned} &\frac{rf(r,\epsilon)}{g(r)} \mathrm{P}\left(\sup_{t\in[0,T]} d\left(\widetilde{q}^r(t), \mathcal{F}\right) > \delta \right) \\ &\qquad \le \frac{rf(r,\epsilon)}{g(r)} \mathrm{P}\left(\sup_{t\in[0,T]} \frac{1}{\theta_{\min}} d\left(\widetilde{q}^r(t), \widetilde{\mathcal{F}}\right) > \delta \right) \underset{r\to\infty}{\longrightarrow} 0, \end{aligned} \tag{83}$$

and part (b) of the theorem follows.

Proof of Part (c). Throughout this proof, we assume that we have fixed a network operated under an MW policy as well as functions $f(\cdot)$ and $g(\cdot)$ with the properties in the statement of the result, namely $\lim_{r\to\infty} rf(r,\epsilon)/g(r) = 0$ for all $\epsilon > 0$ and $\lim_{r\to\infty} g(r)/r = \infty$. It is not hard to see that these properties guarantee that there exists a function $h : \mathbb{N} \to \mathbb{R}_+$ such that

$$\lim_{r\to\infty} \frac{rf(r,\delta)}{h(r)} = 0, \qquad \forall\, \delta > 0, \tag{84}$$

$$\lim_{r\to\infty} \frac{h(r)}{g(r)} = 0, \tag{85}$$

$$\lim_{r\to\infty} \frac{h^2(r)}{rg(r)} = \infty. \tag{86}$$

Let $\widetilde{h}(r) = rg(r)/h(r)$. Then,

$$\lim_{r\to\infty} \frac{\widetilde{h}(r)}{r} = \lim_{r\to\infty} \frac{g(r)}{h(r)} = \infty, \tag{87}$$

$$\lim_{r\to\infty} \frac{\widetilde{h}(r)}{h(r)} = \lim_{r\to\infty} \frac{rg(r)}{h^2(r)} = 0. \tag{88}$$

Before continuing with the main part of the proof, we establish that $\mathcal{F}(\lambda)$ is contained in a low-dimensional subspace. The intuition behind this fact is that $\mathcal{F}(\lambda)$ is contained in the intersection of different effective regions, each of which is a polyhedron. Recall that $\mathcal{C}$ stands for the capacity region of the network.

Claim 4. Suppose that $\lambda \in \mathcal{C}$ but λ is not an extreme point of $\mathcal{C}$. Then, there exists a nonzero vector $v \in \mathbb{R}^n$ such that $v^T\mathcal{F}(\lambda) = \{0\}$.

Proof of Claim 4. Because $\lambda \in \mathcal{C}$, it follows from (5) that $\lambda \in (I - R)\mathrm{Conv}(\mathcal{S})$. Therefore,

$$\lambda = (I - R) \sum_{\mu\in\mathcal{S}} \alpha_\mu \mu, \tag{89}$$

for some nonnegative coefficients α_μ that sum to one. Let us assume that we have fixed one particular set of such coefficients.

Consider some $x \in \mathcal{F}(\lambda)$ and let F be the induced dynamical system of the network. It follows from Proposition 2 and the definition of $\mathcal{F}(\lambda)$ that $0 \in F(x) + \lambda$. Therefore, $\lambda \in -F(x)$. because $F(x)$ is the convex hull of the vectors μ that maximize $x^T(I - R)\mu$, we have

$$\lambda = (I - R) \sum_{\nu\in\mathcal{S}(x)} \beta_\nu \nu, \tag{90}$$

for some nonnegative coefficients β_ν that sum to one. This, together with (89), implies that

$$(I-R)\sum_{\mu\in\mathcal{S}}\alpha_\mu\mu \;=\; (I-R)\sum_{\nu\in\mathcal{S}(x)}\beta_\nu\nu, \tag{91}$$

and as a result,

$$\sum_{\mu\in\mathcal{S}}\alpha_\mu\, x^T(I-R)\mu \;=\; \sum_{\nu\in\mathcal{S}(x)}\beta_\nu\, x^T(I-R)\nu. \tag{92}$$

Because $\mathcal{S}(x)$ is the set of maximizers of $x^T(I-R)\nu$ over $\nu\in\mathcal{S}$, it follows from (92) that, if $\alpha_\mu>0$, then

$$\mu\in\mathcal{S}(x), \tag{93}$$

and this relation is true for all $x\in\mathcal{I}(\lambda)$. This is because, otherwise, the left-hand side of (92) would be strictly smaller than the right-hand side.

On the other hand, because λ is not an extreme point of $\mathcal{C}$, it follows from (5) that λ is not an extreme point of $(I-R)\mathrm{Conv}(\mathcal{S})$. Then, there are at least two service vectors $\mu,\nu\in\mathcal{S}$ for which $\alpha_\mu,\alpha_\nu>0$ and $(I-R)\mu\neq(I-R)\nu$. Let $v=(I-R)(\mu-\nu)$, which is a nonzero vector. As already shown in (93), for any $x\in\mathcal{I}(\lambda)$, we have $\mu,\nu\in\mathcal{S}(x)$. Therefore, for any $x\in\mathcal{I}(\lambda)$, we have $x^T(I-R)\mu=x^T(I-R)\nu$, that is, $v^Tx=0$, and the claim follows. □

Using this claim, consider an $(n-1)$-dimensional subspace Z containing $\mathcal{I}(\lambda)$ and let w be a vector in $\mathbb{R}^n_+\backslash Z$. By suitably scaling w, we can assume that $d(w,Z)=1$. Then, w can be decomposed as $w=z+v$ for some $z\in Z$ and some $v\neq 0$ that is orthogonal to Z and has unit norm $v=1$. We also let

$$b\triangleq\max_{\mu\in\mathcal{S}}\sup_{Q\in\mathbb{R}^n_+}\big\|(I-R)\min\big(\mu,Q\big)\big\|, \tag{94}$$

which is the maximum instantaneous change in the queue lengths resulting from service, where R and $\mathcal{S}$ are the routing matrix and the set of service vectors, respectively. It follows from (87) and (88) that there exists some $r_0\in\mathbb{N}$ such that, for any $r\geq r_0$,

$$2r\delta+\|\lambda\|+b\;<\;\widetilde{h}(r). \tag{95}$$

For every $r\in\mathbb{N}$, let $A^r(\cdot)$ be an i.i.d. process with values

$$A^r(t)=\begin{cases}\lambda+\widetilde{h}(r)\,w, & \text{w.p.}\quad \frac{1}{h(r)},\\ \lambda, & \text{w.p.}\quad 1-\frac{1}{h(r)}.\end{cases} \tag{96}$$

Because $w\in\mathbb{R}^n_+$, $A^r(t)$ is nonnegative for all r and t, and from (88),

$$\mathbb{E}\{A^r(t)\}=\lambda+\frac{\widetilde{h}(r)}{h(r)}w\;\underset{r\to\infty}{\longrightarrow}\;\lambda. \tag{97}$$

For any $r\in\mathbb{N}$ and any $T\in\mathbb{Z}_+$, consider an event E^r_T:

$$E^r_T:\text{the event that } A^r(t)=\lambda+\widetilde{h}(r)\,w,\text{for at least one } t\in[0,T).$$

Consider some $r\geq r_0$. If E^r_r does not occur, then, for any $t<r$ and any $\delta>0$,

$$\left\|\frac{1}{r}\sum_{\tau=0}^{t}(A^r(\tau)-\lambda)\right\|\;=\;0\;<\;\delta.$$

Therefore,

$$\mathrm{P}\left(\frac{1}{r}\sup_{t<r}\left\|\sum_{\tau=0}^{t}(A^r(\tau)-\lambda)\right\|>\delta\right)\;\leq\;\mathrm{P}\big(E^r_r\big)\;\leq\;\frac{r}{h(r)},$$

where the second inequality follows from (96) and the union bound. Thus, for any $\delta>0$,

$$\lim_{r\to\infty}f(r,\delta)\,\mathrm{P}\left(\frac{1}{r}\sup_{t<r}\left\|\sum_{\tau=0}^{t}(A^r(\tau)-\lambda^r)\right\|>\delta\right)\;\leq\;\lim_{r\to\infty}f(r,\delta)\,\frac{r}{h(r)}=0,$$

where the equality is due to (84). Hence, $A^r(\cdot)$, $r\in\mathbb{N}$, is an f-tailed sequence of processes.

According to our definition of v, we have $v = w - z$, v is orthogonal to the subspace Z containing $\mathcal{I}(\lambda)$, and $\|v\| = d(w, Z) = 1$. Therefore, for any $r \geq r_0$, if $A^r(t) = \lambda + \widetilde{h}(r)\, w$ for some t, then

$$\begin{aligned} v^T \left(Q^r(t+1) - Q^r(t)\right) &= v^T A^r(t) + v^T (R - I) \min\big(\mu(t), Q(t)\big) \\ &\geq v^T A^r(t) \; - \; b \|v\| \\ &= \widetilde{h}(r) + v^T \lambda - b \\ &\geq \widetilde{h}(r) - \|\lambda\| - b \\ &> 2r\delta, \end{aligned}$$

where the inequalities are due to (94), $\|v\| = 1$, and (95), respectively. Now, recall that v is orthogonal to the subspace Z containing $\mathcal{I}(\lambda)$. Whenever we have $A^r(t) = \lambda + \widetilde{h}(r)\, v$, we have a jump of size at least $2r\delta$ in a direction orthogonal to $\mathcal{I}(\lambda)$ from which it is not hard to see that

$$\max\left(d(Q^r(t+1),\, \mathcal{I}(\lambda))\,,\, d(Q^r(t),\, \mathcal{I}(\lambda))\right) > \; r\delta. \tag{98}$$

This implies that, for any $r \geq r_0$, if the event $E^r_{g(r)T}$ occurs, then

$$\sup_{t \in [0, g(r)T]} d(Q^r(t)\,,\, \mathcal{I}(\lambda)) > r\delta.$$

Therefore,

$$\begin{aligned} \lim_{r\to\infty} \mathrm{P}\left(\sup_{t\in[0,T]} d\big(\widetilde{q}^r(t)\,,\, \mathcal{I}(\lambda)\big) > \delta \right) &= \lim_{r\to\infty} \mathrm{P}\left(\sup_{t\in[0,g(r)T]} d(Q^r(t)\,,\, \mathcal{I}(\lambda)) > r\delta \right) \\ &\geq \lim_{r\to\infty} \mathrm{P}\left(E^r_{g(r)T}\right) \\ &\geq \lim_{r\to\infty} \left(1 - (1 - 1/h(r))^{g(r)T}\right) \\ &\geq \lim_{r\to\infty} \left(1 - e^{-g(r)T/h(r)}\right) \\ &= 1, \end{aligned}$$

where the last equality is due to (85). This completes the proof of part (c).

Remark 1. The requirement, in part (c) of the theorem, that λ is not an extreme point of $\mathcal{C}$ cannot be removed. For a trivial example, consider a single queue with a single (one-dimensional) service vector $\mu = 1$. Then, $\mathcal{C} = [0, 1]$. If $\lambda = 1$, then every $q \geq 0$ is an invariant state: $\mathcal{I}(1) = [0, \infty)$. The conclusion of Claim 4 fails to hold, and it is certainly impossible for the state to be outside $\mathcal{I}(1)$.

Remark 2. The process $A^r(t)$, used in the proof of part (c) is not uniformly bounded as it can have bursts of size $\widetilde{h}(r)$. With some additional effort and a slightly more complicated proof, it is possible to carry out a construction under which each component of $A^r(t)$ is bounded by some constant, independent of r or t. The basic idea is that having excess arrivals (but of bounded size) over a time period of length $O(\widetilde{h}(r))$ has an effect comparable to a single burst of size $O(\widetilde{h}(r))$.

7. Discussion

In this section, we review our main results, their implications, and directions for future research.

7.1. Main Results

We have established a deterministic bound on the sensitivity of queue length processes with respect to arrivals under a max-weight policy. In particular, we showed that the distance between a queue length process and a fluid solution remains bounded by a constant multiple of the deviation of the aggregate arrival process from its average. The bound allows for tight approximations of the queue lengths in terms of fluid solutions, which are much easier to analyze, and leads to a simple derivation of a fluid limit result; see Corollary 1. We then exploited this sensitivity result to prove matching upper and lower bounds for the time scale over which

additive SSC occurs under a MW policy. As a corollary, we established strong (additive) SSC of MW dynamics in diffusion scaling under conditions more general than previously available.

For the case of i.i.d. arrivals, we established additive SSC over time intervals whose length scales exponentially with r. Such a result could also be proved with a more elementary argument by viewing the distance from the invariant set as a Lyapunov function and using the drift properties that we established; see Lemma 2. Nonetheless, such Lyapunov-based approaches are hard to generalize to broader classes of arrival processes. In contrast, our sensitivity bounds in Theorem 2 allow for the arrival processes to be arbitrary and yield strong approximation results as long as the driving process has some reasonable concentration properties.

7.2. Other Applications and Extensions

A similar sensitivity bound can also be proved, using the same line of argument for continuous time networks, for example, with Poisson arrivals, operating under an MW policy. Similarly, for a more general class of stochastic processing networks, and under assumption 1 of Dai and Lin (2005), it is not hard to see that the fluid dynamics again are a subgradient flow and that our results can be extended to such systems.

In the same spirit, we believe that the results, including additive SSC, can be extended to the case of back-pressure policies[14] (Tassiulas and Ephremides 1992, Georgiadis et al. 2006, Neely 2010) for networks in which the routing is no longer fixed and in which the different vectors μ determine the set of links to be activated.

Another direction concerns SSC results in steady state, that is, the extent to which the steady-state distribution is concentrated in a neighborhood of the invariant set. Following a Lyapunov-based approach, several works (Maguluri and Srikant 2015, Maguluri et al. 2016, Shah et al. 2016) have proved exponential tail bounds for the steady-state distribution for the case of i.i.d. arrivals. We believe that Theorem 2 provides an approach for establishing similar bounds for the case of non-i.i.d. and non-Markovian arrivals.

Finally, another problem in which the fluid model turns out to be analytically beneficial concerns delay stability under an MW policy in the presence of heavy-tailed traffic. Markakis et al. (2016, 2018) studied the question of whether a certain queue has finite expected delay ("delay stability") in the presence of other queues that are faced with heavy-tailed arrivals. They provided a necessary condition for this to be the case in terms of certain properties of the associated fluid model and raised the question of whether, under some assumptions, this condition is also sufficient. Using the sensitivity results of the current paper, we are able to resolve a variant of this question as will be reported in a forthcoming publication.

7.3. Open Problems

The sensitivity bound in Theorem 2 applies only to MW and WMW policies. It is not clear whether a similar bound holds for the more general classes of MW-α and MW-f policies. A similar question also arises about SSC: does Theorem 3 hold under a MW-α policy?

Acknowledgments

This work was partially done while A. Sharifnassab was a visiting student at the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge MA, 02139.

Appendix A. Proof of Lemma 2

In this appendix, we present the proof of Lemma 2. The high-level idea is that $q(\cdot)$ is a trajectory of a subgradient dynamical system corresponding to a convex function that has $\mathcal{G}(\lambda)$ as its set of minimizers. We then show that, in such a system, all trajectories are attracted to the set of minimizers at a uniform rate.

Consider the convex function Φ in (36),

$$\Phi(x) = \max_{\mu \in \mathcal{S}} \big((I-R)\mu\big)^T x,$$

and let $F = -\partial\Phi$ be its subgradient field. Then, F is the induced FPCS system of the network (see Definition 5), and Proposition 2 states that fluid solutions are the same as the nonnegative unperturbed trajectories of $\dot{q} \in F(q) + \lambda$.

We define another convex function Φ_λ by $\Phi_\lambda(x) = \Phi(x) - \lambda^T x$ for all $x \in \mathbb{R}^n$. Because λ is in the capacity region, (5) implies that $\lambda \in (I-R)\mathrm{Conv}(\mathcal{S})$. Then, for any $x \in \mathbb{R}^n$,

$$\begin{aligned} \Phi_\lambda(x) &= \Phi(x) - \lambda^T x \\ &= \max_{\mu \in \mathcal{S}} \big((I-R)\mu\big)^T x - \lambda^T x \\ &\geq \lambda^T x - \lambda^T x \\ &= 0. \end{aligned} \tag{A.1}$$

On the other hand, we have $\Phi_\lambda(0) = 0$. It follows that the set of minimizers of Φ_λ, denoted by Γ, is

$$\Gamma = \{x \in \mathbb{R}^n \,|\, \Phi_\lambda(x) = 0\}. \tag{A.2}$$

We use the shorthand notation $\mathcal{I}$ for $\mathcal{I}(\lambda)$. We now develop a characterization and a simple polyhedral description of the set $\mathcal{I}$. Recall that $\mathcal{I}$ is the set of all $x_0 \in \mathbb{R}^n_+$ for which the constant trajectory $q(t) = x_0$ is a fluid solution. As pointed out earlier, fluid solutions are the same as the nonnegative unperturbed trajectories of the system $F(\cdot) + \lambda$, where $F + \lambda = -\partial\Phi_\lambda$. In particular, $q(t) = x_0$ is a constant trajectory of this system if and only if $\partial\Phi_\lambda(x_0)$ contains the zero vector, which is the case if and only if x_0 is a minimizer of Φ_λ, that is, $x_0 \in \Gamma$. Therefore,

$$\mathcal{I} = \Gamma \cap \mathbb{R}^n_+. \tag{A.3}$$

For any $\mu \in \mathcal{S}$, we define the half space $H_\mu = \{x \in \mathbb{R}_n \,|\, ((I - R)\mu)^T x \le \lambda^T x\}$. We also let, for $i = 1, \ldots, n$, $H_i = \{x \in \mathbb{R}_n \,|\, x_i \ge 0\}$. We now show that

$$\mathcal{I} = \left(\bigcap_{\mu\in\mathcal{S}} H_\mu\right) \bigcap \left(\bigcap_{i=1}^{n} H_i\right). \tag{A.4}$$

Suppose that $x_0 \in \mathcal{I}$. Then, $x_0 \in \mathbb{R}^n_+$, that is, $x_0 \in H_i$ for all i. Furthermore, $x_0 \in \Gamma$ and, from (A.2), $\Phi_\lambda(x_0) = 0$. From (A.1), this implies that $\max_{\mu\in\mathcal{S}}((I - R)\mu^T)x_0 = \lambda^T x_0$ so that $((I - R)\mu)^T x_0 \le \lambda^T x_0$ or, equivalently, $x_0 \in H_\mu$, for all μ. This argument can be reversed. If x_0 belongs to all of the half spaces H_μ, we have $\Phi_\lambda(x_0) \le 0$, which, in light of (A.1) implies that $\Phi_\lambda(x_0) = 0$ or $x_0 \in \Gamma$. If, in addition, $x_0 \in \mathbb{R}^n_+$, then, from (A.3), we obtain $x_0 \in \mathcal{I}$. This concludes the proof of (A.4).

Having characterized the set $\mathcal{I}$, we now turn our attention to the dynamics that drive trajectories toward $\mathcal{I}$. Fix some $\mu \in \mathcal{S}$. For any $x \notin H_\mu$, the closest point to x in H_μ lies on the boundary of H_μ, that is, on the subspace $W_\mu = \{z \in \mathbb{R}^n \,|\, g_\mu^T z = 0\}$, where $g_\mu = (I - R)\mu - \lambda$. For any $x \in \mathbb{R}^n$ and $\mu \in \mathcal{S}$, either $d(x, H_\mu) = 0$ (which includes the case in which $g_\mu = 0$ so that $W_\mu = H_\mu = \mathbb{R}^n$) or $g_\mu \ne 0$ and $x \notin H_\mu$, in which case

$$d(x, H_\mu) = d(x, W_\mu) = \frac{g_\mu^T x}{\|g_\mu\|}, \tag{A.5}$$

which is the length of the projection of x on the normal vector to W_μ. Thus, in both cases, we have

$$\|g_\mu\| \cdot d(x, H_\mu) = \max\left(g_\mu^T x, 0\right). \tag{A.6}$$

Then, for any $x \in \mathbb{R}^n$,

$$\begin{aligned}
\Phi_\lambda(x) &= \max_{\mu\in\mathcal{S}} \left((I - R)\mu - \lambda\right)^T x \\
&= \max_{\mu\in\mathcal{S}} g_\mu^T x \\
&= \max_{\mu\in\mathcal{S}} \max\left(g_\mu^T x, 0\right) \\
&= \max_{\mu\in\mathcal{S}} \|g_\mu\| \cdot d(x, H_\mu).
\end{aligned}$$

Let $\epsilon = \min\{\|g_\mu\| : \mu \in \mathcal{S}, g_\mu \ne 0\}$. Without loss of generality, we assume that $\epsilon > 0$; otherwise, we would be dealing with a trivial system in which every g_μ is zero, and Φ_λ is identically zero. Note that, for any $x \in \mathbb{R}^n_+$, we have

$$\begin{aligned}
\Phi_\lambda(x) &\ge \epsilon \max_{\mu\in\mathcal{S}} d(x, H_\mu) \\
&= \epsilon \max\left(\max_{\mu\in\mathcal{S}} d(x, H_\mu), \max_{i=1,\ldots,n} d(x, H_i)\right),
\end{aligned} \tag{A.7}$$

where the equality is because, when $x \in \mathbb{R}^n_+$, we have $d(x, H_i) = 0$ for all i.

Consider a fluid solution $x(\cdot)$ and fix a time t such that $x(t) \notin \mathcal{F}$. Let $x_0 = x(t)$ and let z be the element of $\mathcal{F}$ that is closest to x_0. Then, at time t,

$$\begin{aligned}
\frac{d}{dt} d(x(t), \mathcal{F}) &= \lim_{h \downarrow 0} \frac{d(x(t+h), \mathcal{F}) - d(x(t), \mathcal{F})}{h} \\
&= \lim_{h \downarrow 0} \frac{d(x(t+h), \mathcal{F}) - d(x(t), z)}{h} \\
&\le \lim_{h \downarrow 0} \frac{d(x(t+h), z) - d(x(t), z)}{h} \\
&= \frac{d^+}{dt} d(x(t), z) \\
&= \frac{(x(t) - z)^T \dot{x}(t)}{\|x(t) - z\|} \\
&= \frac{(x_0 - z)^T \dot{x}(t)}{d(x_0, \mathcal{F})}.
\end{aligned}$$

Putting everything together, we obtain

$$\begin{aligned}
\frac{d}{dt} d(x(t), \mathcal{F}) &\le \frac{(x_0 - z)^T \dot{x}(t)}{d(x_0, \mathcal{F})} \\
&\le \frac{\Phi_\lambda(z) - \Phi_\lambda(x_0)}{d(x_0, \mathcal{F})} \\
&= -\frac{\Phi_\lambda(x_0)}{d(x_0, \mathcal{F})} \\
&\le -\frac{\epsilon \max\big(\max_{\mu \in \mathcal{F}} d\big(x_0, H_\mu\big), \max_{i=1,\ldots,n} d(x_0, H_i)\big)}{d\big(x_0, \big(\bigcap_{\mu \in \mathcal{F}} H_\mu\big) \cap \big(\bigcap_{i=1}^n H_i\big)\big)} \\
&\le -c\,\epsilon,
\end{aligned}$$

where the second inequality is because $-\dot{x}(t)$ is a subgradient of Φ_λ at x_0; the equality is because $z \in \mathcal{F}$ and Φ_λ vanishes on $\mathcal{F}$ by (A.3), (A.2); the third inequality is due to (A.7) and (A.4); and the last inequality follows from Lemma 5 for the constant c therein. This completes the proof of Lemma 2 with $\alpha(\lambda) = c\,\epsilon$.

Appendix B. Proof of Lemma 3

Let $X_1, \ldots, X_t$ be i.i.d. random variables, taking values in $[0, a]$ and let $\overline{X} = (X_1 + \cdots + X_t)/t$. Then, for any $\delta > 0$, Hoeffding's (1963) inequality yields

$$\mathrm{P}\big(\overline{X} - \mathbb{E}\overline{X} \rangle \delta\big) \le 2 \exp\left(-\frac{2t\delta^2}{a^2}\right). \tag{B.1}$$

For any fixed $r \in \mathbb{N}$ and $t \le r$,

$$\begin{aligned}
\mathrm{P}\left(\frac{1}{r}\left\|\sum_{\tau=0}^{t} \big(A^r(\tau) - \lambda^r\big)\right\| > \delta\right) &\le \sum_{i=1}^{n} \mathrm{P}\left(\frac{1}{r}\sum_{\tau=0}^{t} \big(A_i^r(\tau) - \lambda_i^r\big) \rangle \frac{\delta}{\sqrt{n}}\right) \\
&= \sum_{i=1}^{n} \mathrm{P}\left(\frac{1}{t+1}\sum_{\tau=0}^{t} \big(A_i^r(\tau) - \lambda_i^r\big) \rangle \frac{r\delta}{(t+1)\sqrt{n}}\right) \\
&\le 2n \exp\left(-\frac{2(t+1)}{a^2}\left(\frac{r\delta}{(t+1)\sqrt{n}}\right)^2\right) \\
&\le 2n \exp\left(-\left(\frac{2r}{r+1}\right)\frac{r\delta}{na^2}\right),
\end{aligned} \tag{B.2}$$

where the first inequality holds because if the Euclidean norm is above δ, then at least one of the components must be above $\delta/\sqrt{n}$, together with the union bound, and the third inequality is due to (B.1). As in the statement of the lemma, let $\beta \in (0,2)$ and $f(r,\delta) = \exp(\beta r\delta/na^2)$. We then have

$$\begin{aligned} f(r,\delta)\,\mathrm{P}\left(\frac{1}{r}\sup_{t\le r}\left\|\sum_{\tau=0}^{t}(A^r(\tau)-\lambda^r)\right\| > \delta\right) &\le f(r,\delta)\sum_{t\le r}\mathrm{P}\left(\frac{1}{r}\left\|\sum_{\tau=0}^{t}(A^r(\tau)-\lambda^r)\right\| > \delta\right) \\ &\le f(r,\delta)\,2nr\exp\left(-\left(\frac{2r}{r+1}\right)\frac{r\delta}{na^2}\right) \\ &= 2nr\exp\left(\frac{\beta r\delta}{na^2}-\left(\frac{2r}{r+1}\right)\frac{r\delta}{na^2}\right) \underset{r\to\infty}{\longrightarrow} 0, \end{aligned} \tag{B.3}$$

where the second inequality is due to (B.2), and the last implication is because $\beta < 2$.

Endnotes

[1] We note here the important distinction between multiplicative and additive (or strong) state space collapse, which is discussed further in Section 2.4. The literature review here is mostly about multiplicative state space collapse.

[2] For any given $\alpha > 0$, the MW-α policy is an extension of the MW policy in which the "weight" of queue i is proportional to Q_i^α, where Q_i is the length of the queue at node i.

[3] A dynamical system is called nonexpansive if, for any two trajectories, $x(\cdot)$ and $y(\cdot)$, we have $\frac{d}{dt}\|x(t)-y(t)\| \le 0$.

[4] A MW-f policy is obtained by replacing Q^TW in (4) by $f(Q)$, where $f:\mathbb{R}^n \to \mathbb{R}^n$ is a function in an appropriate class.

[5] Max-weight is a special case with the utility function equal to zero.

[6] For a concrete example, if μ corresponds to serving only queue j with unit service rate and if work completed at queue j is routed to queue i, the term $Q^TW(I-R)\mu$ is of the form $w_jQ_j - w_iQ_i$.

[7] This alternative description also explains why uniqueness holds in contrast to the case of more general multiclass queueing networks; see the last paragraph of the proof of Proposition 2.

[8] In our statement of the assumption, we modify the notation of Shah and Wischik (2012), interchanging the roles of z and r to preserve consistency with the rest of this paper.

[9] "Stationary" means that the $A^r(t)$ have the same distribution for all t but without necessarily being independent.

[10] Our rephrasing consists of replacing the term denoted by $\Delta W(\widehat{q}^r(t))$ in Shah and Wischik (2012) by $\mathcal{I}(\lambda)$. This is legitimate because $\Delta W(\widehat{q}^r(t)) \in \mathcal{I}(\lambda)$ (see theorem 5.4 (iv) in Shah and Wischik (2012)), and therefore, $d(\widehat{q}^r(t), \mathcal{I}(\lambda)) \le d(\widehat{q}^r(t), \Delta W(\widehat{q}^r(t)))$.

[11] The result in Shah et al. (2010) assumed that $q_0 = 0$; however, the proof extends to the case of general q_0.

[12] As pointed out by a reviewer, a further extension appears possible, in which q_0 is not an element of $\mathcal{I}(\lambda)$ with the results holding over intervals of the form $[\epsilon, T]$ for any $\epsilon > 0$ in the spirit of similar results in Williams (1998) and Stolyar (2004).

[13] A set-valued function $F:\mathbb{R}^n \to 2^{\mathbb{R}^n}$ is a monotone map if, for any $x_1, x_2 \in \mathbb{R}^n$ and any $v_1 \in F(x_1)$ and $v_2 \in F(x_2)$, we have $(v_1-v_2)^T(x_1-x_2) \le 0$. It is called a maximal monotone map if it is monotone, and for any monotone map $\widetilde{F}$ that satisfies $F(x) \subseteq \widetilde{F}(x)$ for all x, we have $\widetilde{F} = F$.

[14] Back-pressure policies are extensions of the MW policy in which routing is no longer fixed. In particular, there is a fixed set of service vectors, in which each service vector μ associates a rate μ_{ij} to each link ij. A back-pressure policy then chooses at each time a service vector μ that maximizes $\sum_{ij}\mu_{ij}(Q_i - Q_j)$, where the sum is taken over all links ij.

References

Andrews M, Kumaran K, Ramanan K, Stolyar A, Vijayakumar R, Whiting P (2004) Scheduling in a queuing system with asynchronously varying service rates. *Probab. Engrg. Inform. Sci.* 18(2):191–217.

Bramson M (1998) State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems* 30(1–2): 89–140.

Dai J, Lin W (2008) Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Ann. Appl. Probab.* 18(6):2239–2299.

Dai JG, Lin W (2005) Maximum pressure policies in stochastic processing networks. *Oper. Res.* 53(2):197–218.

Dai JG, Prabhakar B (2000) The throughput of data switches with and without speedup. *Proc. INFOCOM 2000*, vol. 2 (IEEE, Piscataway, NJ), 556–564.

Eryilmaz A, Srikant R (2012) Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems* 72(3–4): 311–359.

Eryilmaz A, Srikant R, Perkins JR (2005) Stable scheduling policies for fading wireless channels. *IEEE/ACM Trans. Networking* 13(2):411–424.

Georgiadis L, Neely MJ, Tassiulas L (2006) *Resource Allocation and Cross-Layer Control in Wireless Networks* (Now Publishers, Hanover, MA).

Hoeffding W (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* 58(301):13–30.

Ji T, Athanasopoulou E, Srikant R (2009) Optimal scheduling policies in small generalized switches. *Proc. INFOCOM 2009* (IEEE, Piscataway, NJ), 2921–2925.

Kang W, Williams R (2012) Diffusion approximation for an input-queued switch operating under a maximum weight matching policy. *Stochastic Systems* 2(2):277–321.

Kang W, Kelly F, Lee N, Williams R (2009) State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *Ann. Appl. Probab.* 19(5):1719–1780.

Lin X, Shroff NB, Srikant R (2006) A tutorial on cross-layer optimization in wireless networks. *IEEE J. Selected Areas Comm.* 24(8):1452–1463.

Maguluri ST, Srikant R (2015) Heavy-traffic behavior of the maxweight algorithm in a switch with uniform traffic. *Performance Evaluation Rev.* 43(2):72–74.

Maguluri ST, Srikant R (2016) Heavy traffic queue length behavior in a switch under the maxweight algorithm. *Stochastic Systems* 6(1):211–250.

Maguluri ST, Burle SK, Srikant R (2016) Optimal heavy-traffic queue length scaling in an incompletely saturated switch. *ACM SIGMETRICS* 44(1):13–24.

Maguluri ST, Srikant R, Ying L (2014) Heavy traffic optimal resource allocation algorithms for cloud computing clusters. *Performance Evaluation* 81:20–39.

Mannor S, Tsitsiklis JN (2005) On the empirical state-action frequencies in Markov decision processes under general policies. *Math. Oper. Res.* 30(3):545–561.

Markakis MG, Modiano E, Tsitsiklis JN (2016) Delay stability of back-pressure policies in the presence of heavy-tailed traffic. *IEEE/ACM Trans. Networking* 24(4):2046–2059.

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.

Neely MJ (2010) *Stochastic Network Optimization with Application to Communication and Queueing Systems,* Synthesis Lectures on Communication Networks, vol. 3 (Morgan & Claypool Publishers, Williston, VT).

Reiman MI (1984) Some diffusion approximations with state space collapse. Baccelli F, Fayolle G, eds. *Modelling and Performance Evaluation Methodology* (Springer, Berlin, Heidelberg), 207–240.

Rockafellar R (1970) On the maximal monotonicity of subdifferential mappings. *Pacific J. Math.* 33(1):209–216.

Shah D, Wischik D (2006) Optimal scheduling algorithms for input-queued switches. *Proc. INFOCOM 2006* (IEEE, Piscataway, NJ), 1–11.

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.

Shah D, Tsitsiklis JN, Zhong Y (2010) Qualitative properties of α-weighted scheduling policies. *Performance Evaluation Rev.* 38(1):239–250.

Shah D, Tsitsiklis JN, Zhong Y (2016) On queue-size scaling for input-queued switches. *Stochastic Systems* 6(1):1–25.

Shakkottai S, Srikant R, Stolyar AL (2004) Pathwise optimality of the exponential scheduling rule for wireless channels. *Adv. Appl. Probab.* 34(6):1021–1045.

Sharifnassab A, Tsitsiklis JN, Golestani J (2020) Sensitivity to cumulative perturbations for a class of piecewise constant hybrid systems. *IEEE Trans. Automatic Control*. Forthcoming.

Stewart DE (2011) *Dynamics with Inequalities: Impacts and Hard Constraints* (SIAM, Philadelphia).

Stolyar AL (2004) Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Ann. Appl. Probab.* 14(1):1–53.

Stolyar AL (2005) Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm. *Queueing Systems* 50(4):401–457.

Subramanian VG (2010) Large deviations of max-weight scheduling policies on convex rate regions. *Math. Oper. Res.* 35(4):881–910.

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.

Wang W, Maguluri ST, Srikant R, Ying L (2018) Heavy-traffic delay insensitivity in connection-level models of data transfer with proportionally fair bandwidth sharing. *Performance Evaluation Rev.* 45(2):232–245.

Williams RJ (1998) Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Systems* 30(1–2):27–88.

Xie Q, Lu Y (2015) Priority algorithm for near-data scheduling: Throughput and heavy-traffic optimality. *Proc. INFOCOM 2015* (IEEE, Piscataway, NJ), 963–972.