



# MIT Open Access Articles

## *Stability of Fluid Queueing Systems with Parallel Servers and Stochastic Capacities*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

<b>Citation</b>	Jin, Li, and Saurabh Amin. "Stability of Fluid Queueing Systems with Parallel Servers and Stochastic Capacities." IEEE Transactions on Automatic Control (2018): 1–1.
<b>As Published</b>	<a href="http://dx.doi.org/10.1109/TAC.2018.2808044">http://dx.doi.org/10.1109/TAC.2018.2808044</a>
<b>Publisher</b>	Institute of Electrical and Electronics Engineers (IEEE)
<b>Version</b>	Author's final manuscript
<b>Citable link</b>	<a href="http://hdl.handle.net/1721.1/117403">http://hdl.handle.net/1721.1/117403</a>
<b>Terms of Use</b>	Creative Commons Attribution-Noncommercial-Share Alike
<b>Detailed Terms</b>	<a href="http://creativecommons.org/licenses/by-nc-sa/4.0/">http://creativecommons.org/licenses/by-nc-sa/4.0/</a>

# Stability of Fluid Queueing Systems with Parallel Servers and Stochastic Capacities

Li Jin and Saurabh Amin\*

February 19, 2018

## Abstract

This note introduces a piecewise-deterministic queueing (PDQ) model to study the stability of traffic queues in parallel-link transportation systems facing stochastic capacity fluctuations. The saturation rate (capacity) of the PDQ model switches between a finite set of modes according to a Markov chain, and link inflows are controlled by a state-feedback policy. A PDQ system is stable only if a lower bound on the time-average link inflows does not exceed the corresponding time-average saturation rate. Furthermore, a PDQ system is stable if the following two conditions hold: the nominal mode's saturation rate is high enough that all queues vanish in this mode, and a bilinear matrix inequality (BMI) involving an underestimate of the discharge rates of the PDQ in individual modes is feasible. The stability conditions can be strengthened for two-mode PDQs. These results can be used for design of routing policies that guarantee stability of traffic queues under stochastic capacity fluctuations.

**Index terms:** Traffic control, queueing systems, stability analysis, stochastic switching systems.

## 1 Introduction

Capacity fluctuations in transportation systems can cause significant efficiency losses to the system operators [1]. In practice, these fluctuations can be frequent and also hard to predict deterministically [2, 3]. Thus, traffic control strategies that assume fixed (or nominal) link capacities may fail to limit the inefficiencies resulting from capacity fluctuations, especially when their intensity and/or frequency is non-negligible. In this note, we introduce a simple fluid queueing model with parallel links (servers) that accounts for stochastically varying capacities of individual links, and investigate its stability under a class of feedback control policies. Our analysis is based on known results on stability of continuous-time Markov processes [4, 5], and properties of piecewise-deterministic Markov processes (PDMPs) [6, 7].

Current literature on control of transportation systems with uncertain link or node capacities either assumes a static (but uncertain) capacity model, or considers time-varying capacities [2, 8, 9]. In the former class of models, the actual capacity is assumed to lie in a known set [9], or is realized according to a given probability distribution [10]. Such models are useful for evaluating the system’s performance against worst-case disturbances. The latter class of models is motivated by situations where the capacity is inherently dynamic. These models can enable more accurate assessment of system performance in comparison to static models. In contrast to the above two classes, the model considered in this note is applicable to situations where the capacity can be modeled as a Markovian process [2, 11, 12].

Since we focus on the behavior of aggregate traffic flows, fluid queueing models are better suited to our objectives than the conventional queueing models (e.g.  $M/M/1$ ) [2, 13]. Single server fluid queueing systems with stochastically switching saturation rates have been studied previously; see [8, 14, 15]. This line of work focuses on the analysis of the stationary distribution of queue length under a fixed inflow or an open-loop control policy. Some results are also available on feedback-controlled fluid queueing systems with stochastic capacities [2, 16]. However, to the best of our knowledge, stability of parallel-link fluid queueing systems with uncertain capacities has not been considered before.

In Section 2, we introduce the parallel-link fluid queueing model with a stochastically switching saturation rate vector. This model is called the *piecewise-deterministic queueing* (PDQ) model, since the saturation rate vector switches between a finite set of values, or modes, according to a Markov chain, while the evolution of queue lengths between mode switches is deterministic. Thus, our model belongs to the class of PDMPs [6]. An advantage of this model is that it can be easily calibrated using commonly available traffic data [3]. Furthermore, since the capacity and the queue lengths can be obtained using modern sensing technologies, this model can be used to design capacity-aware control policies.

Our stability notion follows [4, 7] in that a PDQ system is stable if the joint distribution of its state (mode and queue lengths) converges to a unique invariant probability measure. Our analysis involves two assumptions: (i) the mode transition process is ergodic, and (ii) the feedback control policy is bounded and continuous in the queue lengths, and also satisfies a monotonicity condition to ensure that more traffic is routed through links with smaller queues.

Under the above-mentioned assumptions, in Section 3, we derive a necessary condition (Theorem 1) and a sufficient condition (Theorem 2) for stability. The necessary condition is that, for every queue, a suitably defined lower bound on the time-average inflow does not exceed the corresponding link’s time-average saturation rate. The sufficiency result requires two conditions: (i) all queues eventually vanish in a “nominal” mode; and (ii) a lower bound on the discharge rate of the system in individual modes verify a bilinear matrix inequality (BMI). Condition (i) essentially ensures the uniqueness of the invariant measure, and Condition (ii) sets a lower bound on the total flow discharged from the system. Theorem 2 also provides an exponential convergence rate towards the invariant measure. The sufficient conditions for stability can be verified in a more straightforward manner in the case when the PDQ system has two modes (Proposition 1). Furthermore,

under a mode-responsive control policy the necessary and sufficient conditions for a two-mode PDQ coincide (Proposition 2).

Finally, in Section 4, we illustrate some applications of our results for designing stabilizing traffic routing policies in parallel-link networks with stochastic capacity fluctuations.

## 2 Piecewise-Deterministic Queueing System

Consider the PDQ system in Figure 1 (left). A constant *demand*  $A \geq 0$  of traffic arrives at the system and is allocated to  $n$  parallel servers. The *inflow* vector  $F(t) = [F_1(t), \dots, F_n(t)]^T \in \mathbb{R}_{\geq 0}^n$  is such that  $\sum_{k=1}^n F_k(t) = A$  for all  $t \geq 0$ . Traffic can be temporarily stored in queueing buffers and discharged downstream. We denote the vector of *queue lengths* by  $Q(t) = [Q_1(t), \dots, Q_n(t)]^T$ . Let  $U(t) = [U_1(t), \dots, U_n(t)]^T$  denote the vector of stochastic *saturation rates*, where  $U_k(t)$  is the maximum rate at which the  $k$ -th server can release traffic at time  $t$ .

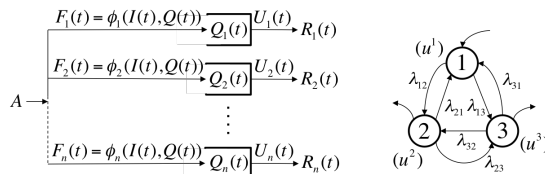


Figure 1: Illustration of a PDQ system with  $n$  parallel servers (left) and the mode transition process (right).

For the  $k$ -th server, if  $Q_k(t) = 0$  and  $F_k(t) \leq U_k(t)$ , the *discharge rate*  $R_k(t)$ , i.e. the rate at which traffic departs from the system through the  $k$ -th server, is given by  $R_k(t) = F_k(t)$ ; otherwise  $R_k(t) = U_k(t)$ . We assume infinite buffer sizes; i.e.  $Q(t)$  can take value in the set  $\mathcal{Q} := \mathbb{R}_{\geq 0}^n$ . This assumption enables us to account for all traffic arriving at the system and not just the traffic that is ultimately discharged by the system.

In our model, the saturation rates of the  $n$  servers stochastically switch between a finite set of values. To model this switching process, we introduce the set of *modes*  $\mathcal{I}$  of the PDQ system and let  $m = |\mathcal{I}|$ . We denote the mode of the PDQ system at time  $t$  by  $I(t)$ . Each mode  $i \in \mathcal{I}$  is associated with a fixed saturation rate, denoted by  $u^i = [u_1^i, \dots, u_n^i]^T$ , which is distinct for each mode. The evolution of  $I(t)$  is governed by a finite-state Markov process with state space  $\mathcal{I}$  and constant transition rates  $\{\lambda_{ij}; i, j \in \mathcal{I}\}$ . We assume that  $\lambda_{ii} = 0$  for all  $i \in \mathcal{I}$ . Note that this is without loss of generality, since self-transitions do not change the saturation rate; thus, including them will not affect the PDQ dynamics. Let

$$\nu_i := \sum_{j \in \mathcal{I}} \lambda_{ij}, \quad (1)$$

which is the rate at which the system leaves mode  $i$ . Given a fixed initial mode  $I_0 \in \mathcal{I}$  at  $t = T_0 := 0$ , let  $\{T_z; z = 1, 2, \dots\}$  be the *epochs* at which the mode transitions occur. Let

$I_{z-1}$  be the mode during  $[T_{z-1}, T_z)$  and  $S_z := T_z - T_{z-1}$ . Then,  $S_z$  follows an exponential distribution with the cumulative distribution function (CDF):

$$F_{S_z}(s) = 1 - e^{-\nu_{I_{z-1}} s}, \quad z = 1, 2, \dots \quad (2)$$

One can capture the transition rates in the  $m \times m$  matrix:

$$\Lambda := \begin{bmatrix} -\nu_1 & \lambda_{12} & \dots & \lambda_{1m} \\ \lambda_{21} & -\nu_2 & \dots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \dots & -\nu_m \end{bmatrix}. \quad (3)$$

We assume the following about the mode transition process:

**Assumption 1.** *The Markov process  $\{I(t); t \geq 0\}$  is ergodic.*

This assumption ensures that the process  $\{I(t); t \geq 0\}$  converges to a unique steady-state distribution, i.e. a row vector  $\mathbf{p} = [p_1, \dots, p_m]$  satisfying the following:

$$\mathbf{p}\Lambda = 0, \quad |\mathbf{p}| = 1, \quad \mathbf{p} \geq 0, \quad (4)$$

where  $|\cdot|$  is the 1-norm.

We consider that the demand  $A$  is distributed across the  $n$  servers according to a state-feedback *routing policy*, which we denote as  $\phi : \mathcal{I} \times \mathcal{Q} \rightarrow \mathbb{R}_{\geq 0}^n$ . A routing policy is *admissible* if  $|\phi(i, q)| = A$  for all  $(i, q) \in \mathcal{I} \times \mathcal{Q}$ . For a given routing policy  $\phi$ , the vector of discharge rates  $R(t)$  is specified by the vector-valued function  $r^\phi : \mathcal{I} \times \mathcal{Q} \rightarrow \mathbb{R}_{\geq 0}^n$  with following components:

$$r_k^\phi(i, q) := \begin{cases} \phi_k(i, q), & q = 0, \phi_k(i, q) \leq u_k^i, \\ u_k^i, & \text{o.w.} \end{cases} \quad k \in \{1, \dots, n\}; \quad (5)$$

i.e., for each  $t \geq 0$ , we have  $R_k(t) = r_k^\phi(I(t), Q(t)) \leq U_k(t)$ .

Let us define a vector field  $D^\phi : \mathcal{I} \times \mathcal{Q} \rightarrow \mathbb{R}^n$  as follows:

$$D^\phi(i, q) := \phi(i, q) - r^\phi(i, q). \quad (6)$$

Then, the evolution of the *hybrid state*  $(I(t), Q(t))$  of the PDQ system is specified by the following dynamics:

$$I(0) = i, \quad Q(0) = q, \quad (i, q) \in \mathcal{I} \times \mathcal{Q}, \quad (7a)$$

$$\Pr\{I(t + \delta) = j' | I(t) = j\} = \lambda_{jj'} \delta + o(\delta), \quad j' \neq j, \quad (7b)$$

$$\frac{dQ(t)}{dt} = D^\phi(I(t), Q(t)), \quad (7c)$$

where  $\delta$  is an infinitesimal time increment. Henceforth, we consider routing policies that satisfy the following assumption:

**Assumption 2.** The routing policy  $\phi(i, q) = [\phi_1(i, q), \dots, \phi_n(i, q)]^T$  is bounded and continuous in  $q$ . Furthermore, for  $k \in \{1, \dots, n\}$ ,  $\phi_k$  is non-increasing in  $q_k$ , and non-decreasing in  $q_h$  for  $h \neq k$ .

The assumption of boundedness and continuity ensures that the Markov process  $\{(I(t), Q(t)); t \geq 0\}$  is right continuous with left limits (RCLL, or *càdlàg*) [6]. Furthermore, since  $Q(t)$  is not reset after mode transitions,  $Q(t)$  is necessarily continuous in  $t$ . With the RCLL property, following [6, Theorem 5.5], the *infinitesimal generator*  $\mathcal{L}^\phi$  of a PDQ with an admissible routing policy  $\phi$  satisfying Assumption 2 is given by

$$\begin{aligned} \mathcal{L}^\phi g(i, q) &= (D^\phi(i, q))^T \nabla_q g(i, q) \\ &+ \sum_{j \in \mathcal{I}} \lambda_{ij} (g(j, q) - g(i, q)), \quad (i, q) \in \mathcal{I} \times \mathcal{Q}, \end{aligned} \quad (8)$$

where  $g$  is any function on  $\mathcal{I} \times \mathcal{Q}$  smooth in the continuous argument.

The assumption of monotonicity of controlled inflows with respect to queue lengths is practically relevant: more traffic is allocated to servers with smaller queues. In addition, this assumption ensures the existence of the following limits:

$$\varphi_{kh}^i := \lim_{q_h \rightarrow \infty} \phi_k(i, q_h \mathbf{e}_h), \quad h, k \in \{1, \dots, n\}, i \in \mathcal{I}, \quad (9)$$

where  $\mathbf{e}_h$  is the  $n$ -dimensional vector such that the  $h$ -th element is 1 and the others are 0. Particularly, the monotonicity of  $\phi$  also implies that  $\phi_k(i, q) \geq \varphi_{kk}^i$  for all  $k \in \{1, \dots, n\}$  and all  $(i, q) \in \mathcal{I} \times \mathcal{Q}$ .

Many practically relevant routing policies satisfy Assumption 2. Examples include:

1. *Mode-responsive* routing policy:

$$\phi_k^{\text{mod}}(i) = \sum_{j \in \mathcal{I}} \mathbb{1}_{\{j=i\}} \psi_k^j, \quad k \in \{1, \dots, n\}, \quad (10)$$

where  $\mathbb{1}_{\{\cdot\}}$  is the indicator function, and  $\psi_k^i \geq 0$  for  $k \in \{1, \dots, n\}$  and  $i \in \mathcal{I}$ . This policy can be viewed as simple re-direction of traffic during disruptions.

2. *Piecewise-affine* routing policy:

$$\begin{aligned} \phi_k^{\text{pwa}}(q) &= \min \left\{ A, \left( \theta_k - \alpha_{kk} q_k + \sum_{h \neq k} \alpha_{kh} q_h \right)_+ \right\}, \\ &k \in \{1, \dots, n\}, \end{aligned} \quad (11)$$

where  $\alpha_{kh} \geq 0$  for all  $k, h \in \{1, \dots, n\}$  and  $(\cdot)_+$  indicates the positive part. This policy is an example of a *queue-responsive* traffic control policy. Note that  $\theta_k$  can be interpreted as the “nominal” inflow sent to each server when no queue exists throughout the system, and the linear terms  $\alpha_{kh} q_h$  as adjustment to these inflows that accounts for the queue lengths.

3. *Logit* routing policy:

$$\phi_k^{\text{log}}(q) = \frac{A \exp(\gamma_k - \beta_k q_k)}{\sum_{h=1}^n \exp(\gamma_h - \beta_h q_h)}, \quad k \in \{1, \dots, n\}, \quad (12)$$

where  $\beta_k \geq 0$  for  $k \in \{1, \dots, n\}$ . This is a classical model of travelers' route choice. One can interpret  $\beta_k$  as sensitivity parameter that reflects travelers' preference to the queue length in the  $k$ -th server, and  $\gamma_k$  the parameter governing travelers' preference when every server has a zero queue.

Note that the computation of the limiting inflows  $\varphi_{kh}^i$  is rather straightforward for the above-mentioned routing policies (see Section 4).

Next, we introduce the notion of stability. The *transition kernel* [4] of a PDQ at time  $t \geq 0$  is a map  $P_t$  from  $\mathcal{I} \times \mathcal{Q}$  to the set of probability measures on  $\mathcal{I} \times \mathcal{Q}$ . Essentially, for an initial condition  $(i, q) \in \mathcal{I} \times \mathcal{Q}$  and a measurable set  $\mathcal{E} \subseteq \mathcal{I} \times \mathcal{Q}$ , we have

$$P_t((i, q); \mathcal{E}) = \Pr\{(I(t), Q(t)) \in \mathcal{E} | I(0) = i, Q(0) = q\}.$$

One can also consider  $P_t$  as an operator acting on probability measures  $\mu$  on  $\mathcal{I} \times \mathcal{Q}$  via

$$\mu P_t(\mathcal{E}) = \int_{\mathcal{I} \times \mathcal{Q}} P_t((i, q); \mathcal{E}) d\mu. \quad (13)$$

An *invariant probability measure* [4] of a PDQ system with routing policy  $\phi$  is a probability measure  $\mu_\phi$  such that

$$\mu_\phi P_t = \mu_\phi, \quad \forall t \geq 0.$$

**Definition 1** (Stability [5, 7]). *The PDQ system with routing policy  $\phi$  is stable if there exists a probability measure  $\mu_\phi$  on  $\mathcal{I} \times \mathcal{Q}$  such that, for each initial condition  $(i, q) \in \mathcal{I} \times \mathcal{Q}$ ,*

$$\lim_{t \rightarrow \infty} \|P_t((i, q); \cdot) - \mu_\phi(\cdot)\|_{\text{TV}} = 0, \quad \forall (i, q) \in \mathcal{I} \times \mathcal{Q}, \quad (14)$$

where  $\|\cdot\|_{\text{TV}}$  is the total variation distance. Furthermore, the PDQ system is exponentially stable if it is stable and there exist constants  $B > 0$  and  $c > 0$ , and a norm-like function<sup>1</sup>  $W : \mathcal{I} \times \mathcal{Q} \rightarrow [1, \infty)$  such that, for any  $(i, q) \in \mathcal{I} \times \mathcal{Q}$ ,

$$\|P_t((i, q); \cdot) - \mu_\phi(\cdot)\|_{\text{TV}} \leq BW(i, q)e^{-ct}, \quad \forall t \geq 0. \quad (15)$$

Finally, the PDQ system is said to be *unstable* if (14) does not hold.

### 3 Stability of Feedback-Controlled PDQs

In this section, we study the stability of controlled PDQ systems. Our main results are Theorem 1 (a necessary condition for stability) and Theorem 2 (a sufficient condition for stability).

---

<sup>1</sup>Following [4],  $W$  is norm-like if  $W(i, q) \rightarrow \infty$  as  $\|q\| \rightarrow \infty$  for  $i \in \mathcal{I}$ .

**Theorem 1.** *Suppose that a PDQ system with  $n$  parallel servers is subject to a total demand  $A \in \mathbb{R}_{\geq 0}$  and is controlled by an admissible policy  $\phi$ . If the PDQ system is stable, then*

$$\sum_{i \in \mathcal{I}} \mathbf{p}_i \varphi_{kk}^i \leq \sum_{i \in \mathcal{I}} \mathbf{p}_i u_k^i, \quad k \in \{1, \dots, n\}, \quad (16)$$

where  $\mathbf{p}_i$  are given by (4) and  $\varphi_{kk}^i$  are given by (9).

*Proof.* Suppose that the PDQ system is stable.

For each server  $k \in \{1, \dots, n\}$  and for each initial condition  $(i, q) \in \mathcal{I} \times \mathcal{Q}$ , we obtain from (6) and (7c) that, for all  $t \geq 0$ ,

$$Q_k(t) = \int_0^t \left( \phi_k(I(s), Q(s)) - r_k^\phi(I(s), Q(s)) \right) ds + q_k.$$

Since  $\lim_{t \rightarrow \infty} q_k/t = 0$ , we have

$$\begin{aligned} 0 &= \lim_{t \rightarrow \infty} \frac{1}{t} \left( \int_0^t \left( \phi_k(I(s), Q(s)) - r_k^\phi(I(s), Q(s)) \right) ds \right. \\ &\quad \left. + q_k - Q_k(t) \right) \\ &= \lim_{t \rightarrow \infty} \frac{1}{t} \left( \int_0^t \left( \phi_k(I(s), Q(s)) - r_k^\phi(I(s), Q(s)) \right) ds \right. \\ &\quad \left. - Q_k(t) \right). \end{aligned} \quad (17)$$

Since the  $k$ -th queue is stable, for each initial condition  $(i, q) \in \mathcal{I} \times \mathcal{Q}$ ,  $\Pr\{\lim_{t \rightarrow \infty} Q(t) = \infty\} = 0$  (i.e. *non- $\nu$ anescence*, see [4, pp. 524] for details), and we have  $\lim_{t \rightarrow \infty} Q_k(t)/t = 0$  a.s. Hence, we can rewrite (17) as

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \left( \phi_k(I(s), Q(s)) - r_k^\phi(I(s), Q(s)) \right) ds = 0, \quad a.s.$$

Now we can make two observations. First, by Assumption 2 (monotonicity), we have

$$\begin{aligned} \phi_k(I(s), Q(s)) &\geq \phi_k(I(s), Q_k(s) \mathbf{e}_k) \\ &\geq \lim_{q_k \rightarrow \infty} \phi_k(I(s), q_k \mathbf{e}_k) = \varphi_{kk}^{I(s)}, \quad \forall s \geq 0. \end{aligned}$$

Secondly, recall that (5) implies  $r_k^\phi(I(s), Q(s)) \leq U_k(s)$  for  $s \geq 0$ . Thus, we have

$$0 = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \left( \phi_k(I(s), Q(s)) - r_k^\phi(I(s), Q(s)) \right) ds$$



$$\geq \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \left( \varphi_{kk}^{I(s)} - U_k(s) \right) ds. \quad (18)$$

In addition, for every  $i \in \mathcal{I}$ , let  $M_i(t)$  be the amount of time that the PDQ system is in mode  $i$  up to time  $t$ , i.e.:

$$M_i(t) = \int_0^t \mathbb{1}_{\{I(s)=i\}} ds.$$

Then, under Assumption 1, we have

$$\lim_{t \rightarrow \infty} \frac{M_i(t)}{t} = \mathbf{p}_i, \text{ a.s. } \forall i \in \mathcal{I}.$$

Hence,

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \varphi_{kk}^{I(s)} ds &= \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \left( \sum_{i \in \mathcal{I}} \mathbb{1}_{\{I(s)=i\}} \varphi_{kk}^i \right) ds \\ &= \lim_{t \rightarrow \infty} \sum_{i \in \mathcal{I}} \frac{M_i(t)}{t} \varphi_{kk}^i = \sum_{i \in \mathcal{I}} \mathbf{p}_i \varphi_{kk}^i, \text{ a.s.} \end{aligned} \quad (19)$$

Similarly, we can obtain

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t U_k(s) ds = \sum_{i \in \mathcal{I}} \mathbf{p}_i u_k^i, \text{ a.s.} \quad (20)$$

Combining (18)–(20), we obtain (16).  $\square$

Theorem 1 provides a way of identifying unstable control policies. As argued in the proof,  $\varphi_{kk}^i$  is in fact the lower bound for  $\phi_k(i, q)$  for all  $q \in \mathcal{Q}$ . Hence, Theorem 1 essentially states that if the PDQ system is stable, then the (time-average) lower bound of the inflow does not exceed the average saturation rate.

To introduce our next result, we define  $R_{\min} = [R_{\min}^1, \dots, R_{\min}^m]^T$  as follows:

$$R_{\min}^i = \min_k \left( u_k^i + \sum_{h: h \neq k} \min\{u_h^i, \varphi_{hk}^i\} \right), \quad i \in \mathcal{I}. \quad (21)$$

One can interpret  $R_{\min}^i$  as a lower bound on the total discharge rate of the  $n$  servers in mode  $i$  when at least one of the  $n$  servers has a non-zero queue. Our next result uses  $R_{\min}^i$  to provide a sufficient condition for the stability of feedback-controlled PDQ systems.

**Theorem 2.** *Suppose that a PDQ system of  $n$  parallel servers is subject to a total demand  $A \in \mathbb{R}_{\geq 0}$  and is controlled by an admissible policy  $\phi$ . Let the elements of the vector  $R_{\min}$  be as defined in (21). Then, the PDQ system is stable if*

$$\exists i^* \in \mathcal{I}, \forall k \in \{1, \dots, n\}, \quad \phi_k(i^*, 0) < u_k^{i^*}, \quad (22)$$

and if

$$\begin{aligned} \exists a = [a^1, \dots, a^m]^T \in \mathbb{R}_{>0}^m, \exists b > 0, \\ \left( \text{diag}(A\mathbf{e} - R_{\min})b + \Lambda \right) a \leq -\mathbf{e}, \end{aligned} \quad (23)$$

where  $\mathbf{e}$  is the  $m$ -dimensional vector of 1's. Furthermore, under the above conditions, there exists a positive constant  $c = \min_{i \in \mathcal{I}} 1/(2a^i)$  such that, for some  $B > 0$ ,

$$\begin{aligned} \|P_t((i, q); \cdot) - \mu_\phi(\cdot)\|_{\text{TV}} \leq B \left( a^i e^{b|q|} + 1 \right) e^{-ct}, \\ \forall (i, q) \in \mathcal{I} \times \mathcal{Q}, \forall t \geq 0, \end{aligned} \quad (24)$$

where  $\mu_\phi$  is the unique invariant probability measure.

The proof of Theorem 2 is based on a more general result [4, Theorem 6.1], which we recall here in the setting of PDQ systems. To conclude stability of the PDQ system, [4, Theorem 6.1] requires that the following two conditions hold:

(A) For any two initial conditions  $(i, q), (j, \ell) \in \mathcal{I} \times \mathcal{Q}$ , there exist  $\delta > 0$  and  $T > 0$  such that

$$\|P_T((i, q); \cdot) - P_T((j, \ell); \cdot)\|_{\text{TV}} \leq 1 - \delta. \quad (25)$$

(B) There exist a norm-like function  $V : \mathcal{I} \times \mathcal{Q} \rightarrow \mathbb{R}_{\geq 0}$  (called the *Lyapunov function*) and constants  $c > 0$  and  $d < \infty$  such that

$$\mathcal{L}V(i, q) \leq -cV(i, q) + d, \quad \forall (i, q) \in \mathcal{I} \times \mathcal{Q}. \quad (26)$$

Condition (A) is required for the uniqueness of the invariant probability measure [17]. Condition (B) is usually referred to as the *drift condition*, which essentially ensures the existence of invariant probability measures [4, Theorem 4.5].

We are now ready to prove the theorem:

*Proof of Theorem 2.* Suppose that (22) and (23) hold. We verify condition (A) (resp. (B)) using (22) (resp. (23)).

Condition (A):

Consider any initial condition  $(i_0, q_0) \in \mathcal{I} \times \mathcal{Q}$ .

First, Assumption 1 ensures that the Markov process  $\{I(t), Q(t); t \geq 0\}$  recurrently visits the mode  $i^*$ . That is, for any  $X_1 > 0$ , there exists  $\sigma > 0$  such that

$$\Pr\{I(X_1) = i^* | I(0) = i_0, Q(0) = q_0\} = \sigma. \quad (27)$$

Furthermore, we can obtain from (7c) that

$$|Q(X_1)| = \left| q_0 + \int_0^{X_1} \left( A - \sum_{k=1}^n r_k^\phi(I(s), Q(s)) \right) ds \right|$$

$$\begin{aligned}
&\leq |q_0| + \int_0^{X_1} \left| A - \sum_{k=1}^n r_k^\phi(I(s), Q(s)) \right| ds \\
&\leq |q_0| + AX_1.
\end{aligned} \tag{28}$$

Secondly, in mode  $i^*$ , the vector of queue length  $Q(t)$  necessarily converges to  $q^* = 0$ . To see this, consider mode  $i^*$  and any  $q \in \mathcal{Q}$ . For each  $k \in \{1, \dots, n\}$  such that  $q_k = 0$ , by Assumption 2, we have  $\phi_k(i^*, q) \geq \phi_k(i^*, 0)$ , and thus

$$\begin{aligned}
r_k^\phi(i^*, q) &= \min\{u_k^{i^*}, \phi_k(i^*, q)\} \\
&\geq \min\{u_k^{i^*}, \phi_k(i^*, 0)\} = r_k^\phi(i^*, 0).
\end{aligned} \tag{29}$$

Therefore, for each  $q \in \mathcal{Q} \setminus \{0\}$ , we have

$$\begin{aligned}
\sum_{k=1}^n D_k^\phi(i^*, q) &\stackrel{(6)}{=} A - \sum_{k=1}^n r_k^\phi(i^*, q) \\
&= A - \sum_{k:q_k>0} u_k^{i^*} - \sum_{k:q_k=0} r_k^\phi(i^*, q) \\
&\stackrel{(29)}{\leq} A - \sum_{k:q_k>0} u_k^{i^*} - \sum_{k:q_k=0} r_k^\phi(i^*, 0) \\
&\leq A - \min_{k:q_k>0} \left( u_k^{i^*} + \sum_{\substack{h:q_h>0 \\ h \neq k}} r_h^\phi(i^*, 0) \right) - \sum_{k:q_k=0} r_k^\phi(i^*, 0) \\
&\leq A - \min_{k \in \{1, \dots, n\}} \left( u_k^{i^*} + \sum_{h \neq k} r_h^\phi(i^*, 0) \right) \\
&\stackrel{(22)}{<} A - \sum_{k=1}^n r_k^\phi(i^*, 0) \stackrel{(5)(22)}{=} 0.
\end{aligned} \tag{30}$$

One can see from (28) and (30) that there exists

$$X_2 = \frac{|q_0| + AX_1}{A - \min_k \left( u_k^{i^*} + \sum_{h \neq k} r_h^\phi(i^*, q) \right)}$$

such that  $Q(X_1 + X_2) = 0$  if  $I(t) = i^*$  for all  $t \in [X_1, X_1 + X_2]$ . Note that

$$\Pr\{I(t) = i^*; t \in [X_1, X_1 + X_2] | I(X_1) = i^*\} = e^{-\nu_{i^*} X_2}.$$

Thus, we have

$$\Pr\{Q(X_1 + X_2) = 0 | I(0) = i_0, Q(0) = q_0\} \geq \sigma e^{-\nu_{i^*} X_2} > 0,$$

where  $\sigma$  satisfies (27). Hence, we have

$$P_{X_1+X_2}((i_0, q_0), \{(i^*, 0)\}) \geq \sigma e^{-\nu_{i^*} X_2}.$$

Then, for any  $T \geq X_1 + X_2$ , we have

$$P_T((i_0, q_0), \{(i^*, 0)\}) \geq \sigma e^{-\nu_{i^*}(T-X_1)}.$$

Thus, for arbitrary initial conditions  $(i, q)$  and  $(j, \ell)$ , there exist  $\sigma' > 0$ ,  $X'_1 > 0$ ,  $X'_2 > 0$ , and  $T' > 0$  such that

$$\begin{aligned} P_{T'}((i, q), \{(i^*, 0)\}) &\geq \sigma' e^{-\nu_{i^*}(T'-X'_2)}, \\ P_{T'}((j, \ell), \{(i^*, 0)\}) &\geq \sigma' e^{-\nu_{i^*}(T'-X'_2)}, \end{aligned}$$

which verifies (25) with  $T = T'$  and  $\delta = \sigma' e^{-\nu_{i^*}(T'-X'_2)}$ .

Condition (B):

Consider the Lyapunov function

$$V(i, q) = a^i e^{b|q|}, \quad (i, q) \in \mathcal{I} \times \mathcal{Q}, \quad (31)$$

where  $a^1, \dots, a^m$ , and  $b$  are positive constants.

For each server  $k$ , by the definition of  $\varphi_{kh}^i$  (9), there necessarily exists  $L_k < \infty$  such that, for all  $h \neq k$ ,

$$\min \left\{ u_h^i, \phi_h(i, L_k \mathbf{e}_k) \right\} \geq \min \left\{ u_h^i, \varphi_{hk}^i \right\} - \frac{1}{2nb \max_{j \in \mathcal{I}} a^j}. \quad (32)$$

Let  $L = [L_1, \dots, L_n]^T$ . We claim that the constants

$$c := \frac{1}{2 \max_{j \in \mathcal{I}} a^j}, \quad (33a)$$

$$d := \max_{i \in \mathcal{I}} |\mathcal{L}V(i, L) + cV(i, L)|, \quad (33b)$$

verify the drift condition (26). Let us prove this claim.

Plugging the Lyapunov function defined in (31) into the expression of the infinitesimal generator (8), we obtain

$$\begin{aligned} \mathcal{L}V(i, q) &= \left( \sum_{k=1}^n \left( \phi_k(i, q) - r_k^\phi(i, q) \right) a^i b \right. \\ &\quad \left. + \sum_{j \in \mathcal{I}} \lambda_{ij} (a^j - a^i) \right) e^{b|q|}. \end{aligned} \quad (34)$$

Then, to check (26), we need to consider two cases:

*Case I:*  $q \in \{\zeta \in \mathcal{Q} : 0 \leq \zeta \leq L\}$ . Since each such  $q$ 's are bounded,  $V(i, q)$  is also bounded. Hence, we can verify in a rather straightforward manner that, with  $c$  and  $d$  given by (33),  $\mathcal{L}V \leq -cV + d$  for all  $i \in \mathcal{I}$  and  $0 \leq q \leq L$ .

*Case II:*  $q \in \mathcal{Q} \setminus \{\zeta \in \mathcal{Q} : 0 \leq \zeta \leq L\}$ . For each such  $q$ , there necessarily exists a server  $k_1$  such that  $q_{k_1} > L_{k_1}$ . For the  $k_1$ -th server, since  $q_{k_1} > L_{k_1} \geq 0$ , we have

$$r_{k_1}^\phi(i, q) = u_{k_1}^i, \quad \forall i \in \mathcal{I}. \quad (35)$$

For the other servers, i.e. for each  $h \neq k_1$ , we have

$$\begin{aligned} r_h^\phi(i, q) &\stackrel{(5)}{=} \min \left\{ u_h^i, \phi_h(i, q) \right\} \\ &\geq \min \left\{ u_h^i, \phi_h(i, q_{k_1} \mathbf{e}_{k_1}) \right\} \geq \min \left\{ u_h^i, \phi_h(i, L_{k_1} \mathbf{e}_{k_1}) \right\} \end{aligned} \quad (36a)$$

$$\stackrel{(32)}{\geq} \min \left\{ u_h^i, \varphi_h^{k_1}(i) \right\} - \frac{1}{2nb \max_{j \in \mathcal{I}} a^j}, \quad \forall i \in \mathcal{I}, \quad (36b)$$

where (36a) results from Assumption 2 (monotonicity). Combining (35) and (36b), we can write

$$\begin{aligned} \sum_{h=1}^n r_h^\phi(i, q) &\geq u_{k_1}^i + \sum_{h:h \neq k_1} \min \left\{ u_h^i, \varphi_h^{k_1}(i) \right\} - \frac{1}{2b \max_{j \in \mathcal{I}} a^j} \\ &\stackrel{(21)}{\geq} R_{\min}^i - \frac{1}{2b \max_{j \in \mathcal{I}} a^j}. \end{aligned} \quad (37)$$

Then,

$$\begin{aligned} &\sum_{k=1}^n \left( \phi_k(i, q) - r_k^\phi(i, q) \right) a^i b + \sum_{j \in \mathcal{I}} \lambda_{ij} (a^j - a^i) \\ &\stackrel{(37)}{\leq} \left( A - R_{\min}^i + \frac{1}{2b \max_{j \in \mathcal{I}} a^j} \right) a^i b + \sum_{j \in \mathcal{I}} \lambda_{ij} (a^j - a^i) \\ &\stackrel{(23)}{\leq} -1 + \frac{1}{2} = -\frac{1}{2}. \end{aligned}$$

Finally,

$$\mathcal{L}V(i, q) \stackrel{(34)}{\leq} -\frac{1}{2} e^{b|q|} \stackrel{(33a)}{\leq} -c a^i e^{b|q|} \stackrel{(31)}{=} -cV.$$

Hence, (26) holds for all  $i \in \mathcal{I}$ , all  $q \in \mathcal{Q} \setminus \{q : 0 \leq \zeta \leq L\}$ , and all  $d \geq 0$ .

Thus, we have verified that the drift condition (26) holds for all  $(i, q) \in \mathcal{I} \times \mathcal{Q}$ .

Finally, note that we have verified conditions (A) and (B) for the controlled PDQ system. Thus, we obtain from [4, Theorem 6.1] that the PDQ system is exponentially stable.  $\square$

The condition (22) states that there exists a mode  $i^*$  in which every queue decreases to zero. Practically, one can interpret  $i^*$  as a “nominal” or “normal” mode in which the saturation rates are sufficiently high and satisfy (22). This condition leads to Condition (A).

The condition (23) essentially imposes a lower bound on the total discharged flow from the  $n$  servers, which is characterized by  $R_{\min}^i$ . This condition leads to Condition (B). To verify this condition, one needs to determine whether BMI (23) admits positive solutions for  $a_1, \dots, a_m$  and  $b$ . This can be done using the known computational methods to solve BMIs (see e.g. [18, 19]).

**Remark 1.** Using the exponential Lyapunov function (31), one can also apply [4, Theorem 4.3] to obtain that, under (23), for each initial condition  $(i, q) \in \mathcal{I} \times \mathcal{Q}$ , we have

$$\limsup_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mathbb{E}[e^{|\mathcal{Q}(s)|}] ds < \infty.$$

That is, moments of the queue lengths are bounded.

Furthermore, if the system has only two modes, solutions for  $b$  and  $a$  can be constructed in a more straightforward manner, which motivates the next result.

**Proposition 1.** A PDQ system of  $n$  parallel servers with two modes  $\{1, 2\}$  and with an admissible control policy  $\phi$  is stable if

$$\exists i^* \in \{1, 2\}, \phi_k(i^*, 0) < u_k^{i^*}, k \in \{1, 2\}, \quad (38)$$

and if

$$A < \mathbf{p}_1 R_{\min}^1 + \mathbf{p}_2 R_{\min}^2, \quad (39)$$

where  $R_{\min}^i$  is defined in (21).

*Proof.* First, let us define the following quantities

$$D_{\min} = \min \{ A - R_{\min}^1, A - R_{\min}^2 \}, \quad (40a)$$

$$D_{\max} = \max \{ A - R_{\min}^1, A - R_{\min}^2 \}, \quad (40b)$$

$$\bar{D} = A - (\mathbf{p}_1 R_{\min}^1 + \mathbf{p}_2 R_{\min}^2), \quad (40c)$$

$$i_{\min} = \begin{cases} 1, & \text{if } D_{\min} = A - R_{\min}^1, \\ 2, & \text{o.w.} \end{cases} \quad (40d)$$

$$i_{\max} = \begin{cases} 2, & \text{if } D_{\min} = A - R_{\min}^1, \\ 1, & \text{o.w.} \end{cases} \quad (40e)$$

$$\lambda_{\min} = \begin{cases} \lambda_{12}, & \text{if } D_{\min} = A - R_{\min}^1, \\ \lambda_{21}, & \text{o.w.} \end{cases} \quad (40f)$$

$$\lambda_{\max} = \begin{cases} \lambda_{21}, & \text{if } D_{\min} = A - R_{\min}^1, \\ \lambda_{12}, & \text{o.w.} \end{cases} \quad (40g)$$

Under (39), we explicitly construct constants  $a^{i_{\min}}$ ,  $a^{i_{\max}}$ , and  $b$  satisfying the BMI (23). Condition (39) implies

$$A - \mathbf{p}_1 R_{\min}^1 - \mathbf{p}_2 R_{\min}^2 = \mathbf{p}_{i_{\min}} D_{\min} + \mathbf{p}_{i_{\max}} D_{\max} < 0 \quad (41)$$

Since  $D_{\min} \leq D_{\max}$ , (41) implies that  $D_{\min} < 0$ . Thus, we only need to consider two cases:

In the case that  $D_{\min} < 0$ ,  $D_{\max} \leq 0$ , we can select an arbitrary  $a^{i_{\min}} > \max_i \{1/\lambda_i\}$  and let

$$a^{i_{\max}} = 2a^{i_{\min}}, \quad b = \frac{\lambda_{\min} a^{i_{\min}} + 1}{-D_{\min} a^{i_{\min}}}. \quad (42)$$

It is not hard to see that  $a^{i_{\min}}$ ,  $a^{i_{\max}}$ , and  $b$  are positive and satisfy the BMI (23).

In the case that  $D_{\min} < 0$ ,  $D_{\max} > 0$ , we let

$$b = \frac{(\lambda_{12} + \lambda_{21})\bar{D}}{2D_{\min}D_{\max}}, \quad (43a)$$

$$a^{i_{\min}} = \frac{-D_{\max}b + \lambda_{12} + \lambda_{21}}{\det[\text{diag}(A\mathbf{e} - R_{\min})b + \Lambda]}, \quad (43b)$$

$$a^{i_{\max}} = \frac{-D_{\min}b + \lambda_{12} + \lambda_{21}}{\det[\text{diag}(A\mathbf{e} - R_{\min})b + \Lambda]}. \quad (43c)$$

Now, we show that these constants are positive. First, note that (39) implies  $\bar{D} < 0$ . Then, since  $D_{\min} < 0$  and  $D_{\max} > 0$ , and since  $\bar{D} < 0$ ,  $b$  is positive. Secondly, to see that  $a^{i_{\min}} > 0$ , note that

$$\begin{aligned} & \text{diag}(A\mathbf{e} - R_{\min})b + \Lambda \\ &= \begin{bmatrix} b(A - R_{\min}^1) - \lambda_{12} & \lambda_{12} \\ \lambda_{21} & b(A - R_{\min}^2) - \lambda_{21} \end{bmatrix}, \end{aligned}$$

and

$$\begin{aligned} & \det[\text{diag}(A\mathbf{e} - R_{\min})b + \Lambda] \\ &= b^2 (A - R_{\min}^1) (A - R_{\min}^2) \\ & \quad - \lambda_{12}b (A - R_{\min}^2) - \lambda_{21}b (A - R_{\min}^1) \\ &= b^2 (A - R_{\min}^1) (A - R_{\min}^2) - b(\lambda_{12} + \lambda_{21})\bar{D} \\ &= b^2 D_{\min} D_{\max} - b(\lambda_{12} + \lambda_{21})\bar{D}. \end{aligned}$$

Again, since  $D_{\min} < 0$  and  $D_{\max} > 0$ , one can check that the  $b$  given in (43a) ensures that  $\det[\text{diag}(A\mathbf{e} - R_{\min})b + \Lambda] > 0$ . In addition, note that

$$\begin{aligned} b &= \frac{(\lambda_{12} + \lambda_{21})\bar{D}}{2D_{\min}D_{\max}} \\ &= \frac{\lambda_{12} + \lambda_{21}}{D_{\max}} \left( \frac{-\mathbf{p}_{i_{\min}}D_{\min} - \mathbf{p}_{i_{\max}}D_{\max}}{-2D_{\min}} \right) \\ &< \frac{\lambda_{12} + \lambda_{21}}{D_{\max}} \left( \frac{-\mathbf{p}_{i_{\min}}D_{\min} - \mathbf{p}_{i_{\max}}D_{\min}}{-2D_{\min}} \right) \\ &= \frac{\lambda_{12} + \lambda_{21}}{2D_{\max}} < \frac{\lambda_{12} + \lambda_{21}}{D_{\max}}, \end{aligned}$$

which, along with  $D_{\max} > 0$ , implies  $a^{i_{\min}} > 0$ . Finally, since  $D_{\min} < 0$ ,  $a^{i_{\max}}$  is also positive.

From (40d) and (40e), we know that

$$a^1 = \begin{cases} a^{i_{\min}}, & \text{if } D_{\min} = A - R_{\min}^1, \\ a^{i_{\max}}, & \text{o.w.} \end{cases}$$

$$a^2 = \begin{cases} a^{i_{\max}}, & \text{if } D_{\min} = A - R_{\min}^1, \\ a^{i_{\min}}, & \text{o.w.} \end{cases}$$

Let  $a = [a^1, a^2]^T$ . Then, one can check that  $a$  and  $b$  satisfy

$$[\text{diag}(A\mathbf{e} - R_{\min})b + \Lambda]a = -\mathbf{e},$$

and thus satisfy the BMI (23).

In addition, (38) is analogous to (22). Thus, we can conclude from Theorem 2 that the two-mode PDQ system is stable. □

In comparison to Theorem 2, Proposition 1 provides a simpler criterion (39) for stability of PDQ systems with two modes, since it does not involve solving a BMI.

Furthermore, if a PDQ system with two modes is controlled by a mode-responsive routing policy (10), then we can obtain a *necessary and sufficient condition* for stability:

**Proposition 2.** *A system of  $n$  parallel servers two modes  $\{1, 2\}$  and with a mode-responsive routing policy  $\phi$  given by (10) is stable if and only if*

$$\mathbf{p}_1\psi_k^1 + \mathbf{p}_2\psi_k^2 < \mathbf{p}_1u_k^1 + \mathbf{p}_2u_k^2, \quad \forall k \in \{1, \dots, n\}. \quad (44)$$

*Proof.* Since the system is controlled by a mode-responsive policy, the queues in various servers do not interact. Therefore, we can consider the  $n$  servers independently. For the  $k$ -th server, consider the Lyapunov function

$$V_k(i, q_k) = a_k^i \exp(b_k q_k), \quad (i, q) \in \{1, 2\} \times \mathbb{R}_{\geq 0}$$

with parameters  $[a_k^1, a_k^2]^T \in \mathbb{R}_{>0}^2$  and  $b_k > 0$ . With this Lyapunov function, one can adapt the proof of Proposition 1 and conclude that the  $k$ -th server is stable if (44) holds.

To obtain the necessity of (44), first note that the  $k$ -th server is unstable if  $\mathbf{p}_1\psi_k^1 + \mathbf{p}_2\psi_k^2 > \mathbf{p}_1u_k^1 + \mathbf{p}_2u_k^2$ . Secondly, to argue that the  $k$ -th server is unstable if

$$\mathbf{p}_1\psi_k^1 + \mathbf{p}_2\psi_k^2 = \mathbf{p}_1u_k^1 + \mathbf{p}_2u_k^2, \quad (45)$$

one can first assume by contradiction the existence of an invariant probability measure  $\mu_\phi$ , and then consider  $\mu_\phi(\mathcal{I} \times \{0\})$  to arrive at a contradiction to (45). □

In addition, for the setting of Proposition 2, expression for the invariant probability measure  $\mu_\phi$  has been reported in the literature [15], which makes possible analytical optimization of the routing policy.

## 4 Illustrative examples

In this section, we demonstrate how our results can provide insights for traffic flow routing under stochastic capacity fluctuation. Consider a network of two parallel servers. The total inflow is  $A = 1$ . Our results in Section 3 can be applied to obtain stability conditions of this network. We particularly focus on the practically motivated routing policies given in (10)–(12).



## 4.1 A two-mode network

Suppose that the network has two modes  $\{1, 2\}$  with symmetric transition rates  $\lambda_{12} = \lambda_{21} = 1$ . Thus, the steady-state probabilities are  $\mathbf{p}_1 = \mathbf{p}_2 = 0.5$ . The saturation rates in both modes are given as  $u^1 = [1.2, 0.7]^T$  and  $u^2 = [0.2, 0.7]^T$ . Thus, both servers have an average saturation rate of 0.7.

### 4.1.1 Mode-responsive routing

For this two-mode system, the policy given by (10) can be parametrized by two constants  $\psi_1^1, \psi_1^2 \in [0, 1]$  (note that admissibility requires  $\psi_1^i + \psi_2^i = 1$  for  $i \in \{1, 2\}$ ). By Proposition 2, the routing policy  $\phi^{\text{mod}}$  is stabilizing if and only if

$$0.3 < (\psi_1^1 + \psi_1^2)/2 < 0.7, \quad \psi_1^1 \in [0, 1], \psi_1^2 \in [0, 1].$$

That is, the PDQ system is stable if and only if the average inflows into each server are less than their respective average saturation rate (note that  $(\psi_1^1 + \psi_1^2)/2 > 0.3$  is equivalent to  $(\psi_2^1 + \psi_2^2)/2 < 0.7$ ).

### 4.1.2 Piecewise-affine feedback routing

Consider the policy given by (11). Admissibility requires  $\alpha_{11} = \alpha_{12}$ ,  $\alpha_{21} = \alpha_{22}$ , and  $\theta_1 + \theta_2 = 1$ . Hence, we denote  $\alpha_1 = \alpha_{11} = \alpha_{12}$  and  $\alpha_2 = \alpha_{21} = \alpha_{22}$ . For  $k = 1, 2$  and  $i \in \{1, 2\}$ , the expression of the limiting inflows (9) are as follows:

$$\varphi_{kk}^i = \begin{cases} 0, & \text{if } \alpha_k > 0, \\ \min\{A, \theta_k\}, & \text{if } \alpha_k = 0, \end{cases}$$

$$\varphi_{kh}^i = \begin{cases} 1, & \text{if } \alpha_h > 0, \\ \min\{A, \theta_k\}, & \text{if } \alpha_h = 0, \end{cases} \quad h \neq k.$$

Table 1 shows the necessary condition for stability given by Theorem 1 and the sufficient condition for stability given by Proposition 1. Note that the restriction on  $\theta_k$  is stronger

Table 1: Stability conditions (two modes, PWA routing).

$\alpha_1$	$\alpha_2$	Necessary condition	Sufficient condition
= 0	= 0	$0.3 \leq \theta_1 \leq 0.7$	$0.3 < \theta_1 < 0.7$
= 0	> 0	$\theta_1 \leq 0.7$	$0.3 < \theta_1 < 0.7$
> 0	= 0	$\theta_1 \geq 0.3$	$\theta_1 > 0.3$
> 0	> 0	$\theta_1 \in \mathbb{R}$	$\theta_1 > 0.3$

if  $\alpha_k = 0$ . The intuition is that, if the routing policy is not responsive to the queue length in a server, then an appropriate selection of the nominal inflow  $\theta_k$  is crucial to ensure stability. In addition, the structures of the stability conditions strongly depend on whether  $\alpha_k$  is zero, but not on the exact magnitude of  $\alpha_k$ . In this example, the gap between the necessary condition and the sufficient condition mainly results from the condition (22), which requires  $\theta_1 > 0.3$ .

### 4.1.3 Logit routing

Now, consider the policy (12). For  $k = 1, 2$ ,  $i \in \{1, 2\}$ , the limiting inflows are

$$\varphi_{kk}^i = \begin{cases} 0, & \text{if } \beta_k > 0, \\ \frac{A \exp(\gamma_k)}{\sum_{h=1}^2 \exp(\gamma_h)}, & \text{if } \beta_k = 0, \end{cases} \quad (46a)$$

$$\varphi_{kh}^i = \begin{cases} A, & \text{if } \beta_h > 0, \\ \frac{A \exp(\gamma_k)}{\sum_{h=1}^2 \exp(\gamma_h)}, & \text{if } \beta_h = 0, \end{cases} \quad h \neq k. \quad (46b)$$

Again, we can obtain a stability conditions from Theorem 1 and Proposition 1. Table 2

Table 2: Stability conditions (two modes, logit routing).

$\beta_1$	$\beta_2$	Necessary condition	Sufficient condition
$= 0$	$= 0$	$ \gamma_1 - \gamma_2  \leq \log(7/3)$	$ \gamma_1 - \gamma_2  < \log(7/3)$
$= 0$	$> 0$	$\gamma_1 - \gamma_2 \leq \log(7/3)$	
$> 0$	$= 0$	$\gamma_1 - \gamma_2 \geq -\log(7/3)$	
$> 0$	$> 0$	$\gamma_1 \in \mathbb{R}, \gamma_2 \in \mathbb{R}$	

implies that the constants  $\gamma_k$  have a stronger impact on stability of the PDQ system than the coefficients  $\beta_k$  capturing the sensitivity to queue lengths. Once again, the gap between the necessary condition and the sufficient condition results from (22), which requires  $|\gamma_1 - \gamma_2| < \log(7/3)$ .

## 4.2 A three-mode network

Suppose that the network has three modes  $\{1, 2, 3\}$  with symmetric transition rates  $\lambda_{ij} = 1$  for all  $i, j \in \mathcal{I}$ . Thus, the steady-state probabilities are  $\mathbf{p}_1 = \mathbf{p}_2 = \mathbf{p}_3 = 1/3$ . The saturation rates in the three modes are  $u^1 = [1.2, 0.7]^T$ ,  $u^2 = [0.7, 0.7]^T$ , and  $u^3 = [0.2, 0.7]^T$ ; i.e. the average saturation rates are equal to those in the two-mode case. The main difference between the analysis in this subsection and that in the previous subsection is that the sufficient conditions for stability below are obtained numerically (in terms of solving the BMI (23)) instead of analytically.

### 4.2.1 Mode-responsive routing

For ease of presentation, we assume that  $\psi_k^2 = \psi_k^3$  for  $k \in \{1, 2\}$ . The limiting inflows  $\varphi_{kh}^i$  are given by

$$\varphi_{kh}^i = \psi_k^i, \quad h \in \{1, 2\}, k \in \{1, 2\}, i \in \mathcal{I}.$$

Theorem 1 gives a necessary condition for stability:

$$0.3 \leq 1/3\psi_1^1 + 2/3\psi_1^2 \leq 0.7, \quad (47)$$

whose complement is the “Unstable” region in Figure 2. Figure 2 also shows a “Stable” region obtained from Theorem 2; the BMI (23) is solved using YALMIP [19]. In contrast to the two-mode case, there is an “Unknown” region between the “Stable” and “Unstable” regions, due to the gap between the necessary condition (Theorem 1) and the sufficient condition (Theorem 2).

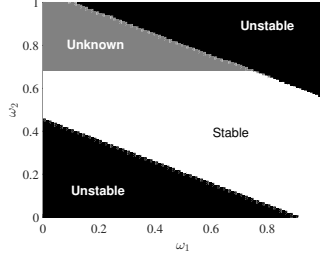


Figure 2: Stability of various  $(\psi_1^1, \psi_1^2)$  pairs.

#### 4.2.2 Queue-responsive routing policies

For the piecewise-affine routing policy (11) and the logit routing policy (12), Tables 3

Table 3: Stability conditions (three modes, PWA routing).

$\alpha_1$	$\alpha_2$	Necessary condition	Sufficient condition
$= 0$	$= 0$	$0.3 \leq \theta_1 \leq 0.7$	$0.41 \leq \theta_1 \leq 0.59$
$= 0$	$> 0$	$\theta_1 \leq 0.7$	$0.41 \leq \theta_1 \leq 0.59$
$> 0$	$= 0$	$\theta_1 \geq 0.3$	$\theta_1 \geq 0.36$
$> 0$	$> 0$	$\theta_1 \in \mathbb{R}$	$\theta_1 > 0.3$

Table 4: Stability conditions (three modes, logit routing).

$\beta_1$	$\beta_2$	Necessary condition	Sufficient condition
$= 0$	$= 0$	$ \gamma_1 - \gamma_2  \leq \log(7/3)$	$ \gamma_1 - \gamma_2  \leq \log 1.7$
$= 0$	$> 0$	$\gamma_1 - \gamma_2 \leq \log(7/3)$	
$> 0$	$= 0$	$\gamma_1 - \gamma_2 \geq -\log(7/3)$	
$> 0$	$> 0$	$\gamma_1 \in \mathbb{R}, \gamma_2 \in \mathbb{R}$	

and 4 show the stability conditions. In comparison to the two-mode case, the necessary conditions are unchanged, but the sufficient conditions in the three-mode case are more restrictive. This indicates that the sufficient condition becomes more restrictive as the number of modes (and thus the number of bilinear inequality constraints) increases.

## Acknowledgments

This work was supported by NSF CNS-1239054 CPS Frontiers, NSF CAREER Award CNS-1453126, and AFRL Lablet-Secure and Resilient Cyber-Physical Systems. We are deeply grateful for the insightful comments from the three anonymous reviewers and the associate editor.

## References

- [1] J. Kwon, M. Mauch, and P. Varaiya, “Components of congestion: Delay from incidents, special events, lane closures, weather, potential ramp metering gain, and excess demand,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1959, no. 1, pp. 84–91, 2006.
- [2] M. D. Peterson, D. J. Bertsimas, and A. R. Odoni, “Models and algorithms for transient queueing congestion at airports,” *Management Science*, vol. 41, no. 8, pp. 1279–1295, 1995.
- [3] L. Jin and S. Amin, “Calibration of a macroscopic traffic flow model with stochastic saturation rates,” in *Transportation Research Board 96th Annual Meeting*, 2017.
- [4] S. P. Meyn and R. L. Tweedie, “Stability of Markovian processes III: Foster-Lyapunov criteria for continuous-time processes,” *Advances in Applied Probability*, pp. 518–548, 1993.
- [5] B. Cloez, M. Hairer, *et al.*, “Exponential ergodicity for Markov processes with random switching,” *Bernoulli*, vol. 21, no. 1, pp. 505–536, 2015.
- [6] M. H. A. Davis, “Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models,” *Journal of the Royal Statistical Society. Series B. Methodological*, vol. 46, no. 3, pp. 353–388, 1984.
- [7] M. Benaïm, S. Le Borgne, F. Malrieu, and P.-A. Zitt, “Qualitative properties of certain piecewise deterministic Markov processes,” in *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, vol. 51, pp. 1040–1075, Institut Henri Poincaré, 2015.
- [8] D. Anick, D. Mitra, and M. M. Sondhi, “Stochastic theory of a data-handling system with multiple sources,” *The Bell System Technical Journal*, vol. 61, no. 8, pp. 1871–1894, 1982.
- [9] G. Como, K. Savla, D. Acemoglu, M. A. Dahleh, and E. Frazzoli, “Robust distributed routing in dynamical networks Part I: Locally responsive policies and weak resilience,” *Automatic Control, IEEE Transactions on*, vol. 58, no. 2, pp. 317–332, 2013.

- [10] G. D. Glockner and G. L. Nemhauser, “A dynamic network flow problem with uncertain arc capacities: Formulation and problem structure,” *Operations Research*, vol. 48, no. 2, pp. 233–242, 2000.
- [11] M. Baykal-Gürsoy, W. Xiao, and K. Ozbay, “Modeling traffic flow interrupted by incidents,” *European Journal of Operational Research*, vol. 195, no. 1, pp. 127–138, 2009.
- [12] L. Jin and S. Amin, “A piecewise-deterministic Markov model of freeway accidents,” in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, IEEE, 2014.
- [13] G. F. Newell, *Applications of Queueing Theory*, vol. 4. Springer Science & Business Media, 2013.
- [14] H. Chen and D. D. Yao, “A fluid model for systems with random disruptions,” *Operations Research*, vol. 40, no. 3-supplement-2, pp. S239–S247, 1992.
- [15] V. G. Kulkarni, “Fluid models for single buffer systems,” *Frontiers in queueing: Models and applications in science and engineering*, vol. 321, p. 338, 1997.
- [16] H. Yu and C. G. Cassandras, “Perturbation analysis of feedback-controlled stochastic flow systems,” *IEEE Transactions on Automatic Control*, vol. 49, no. 8, pp. 1317–1332, 2004.
- [17] J. G. Dai and S. P. Meyn, “Stability and convergence of moments for multiclass queueing networks via fluid limit models,” *IEEE Transactions on Automatic Control*, vol. 40, no. 11, pp. 1889–1904, 1995.
- [18] J. G. VanAntwerp and R. D. Braatz, “A tutorial on linear and bilinear matrix inequalities,” *Journal of process control*, vol. 10, no. 4, pp. 363–385, 2000.
- [19] J. Löfberg, “YALMIP: A toolbox for modeling and optimization in MATLAB,” in *Computer Aided Control Systems Design, 2004 IEEE International Symposium on*, pp. 284–289, IEEE, 2004.