Stochastic Systems
2016, Vol. 6, No. 1, 132–210
DOI: 10.1214/15-SSY187

CHATTERING AND CONGESTION COLLAPSE IN AN OVERLOAD SWITCHING CONTROL

By Ohad Perry[*] and Ward Whitt[†]

Northwestern University[*] and Columbia University[†]

Routing mechanisms for stochastic networks are often designed to produce *state space collapse* (SSC) in a heavy-traffic limit, i.e., to confine the limiting process to a lower-dimensional subset of its full state space. In a fluid limit, a control producing asymptotic SSC corresponds to an ideal *sliding mode* control that forces the fluid trajectories to a lower-dimensional *sliding manifold*. Within deterministic dynamical systems theory, it is well known that sliding-mode controls can cause the system to chatter back and forth along the sliding manifold due to delays in activation of the control. For the prelimit stochastic system, chattering implies fluid-scaled fluctuations that are larger than typical stochastic fluctuations.

In this paper we show that chattering can occur in the fluid limit of a controlled stochastic network when inappropriate control parameters are used. The model has two large service pools operating under the *fixed-queue-ratio with activation and release thresholds* (FQR-ART) overload control which we proposed in a recent paper. The FQR-ART control is designed to produce asymptotic SSC by automatically activating sharing (sending some customers from one class to the other service pool) once an overload occurs. We have previously shown that this control is effective and robust, even if the service rates are less for the other shared customers, when the control parameters are chosen properly. We now show that, if the control parameters are not chosen properly, then delays in activating and releasing the control can cause chattering with large oscillations in the fluid limit. In turn, these fluid-scaled fluctuations lead to severe congestion, even when the arrival rates are smaller than the potential total service rate in the system, a phenomenon referred to as *congestion collapse*. We show that the fluid limit can be a bi-stable switching system possessing a unique nontrivial periodic equilibrium, in addition to a unique stationary point.

1. Introduction. In this paper we study the fluid limit of a stochastic system comprised of two service pools, each having its own arrival process and own queue. The system is operating under the *fixed-queue-ratio with*

Received July 2015.

MSC 2010 subject classifications: Primary 60K25, 34C25, 34C55; secondary 60F1f7, 37G15

Keywords and phrases: Stochastic networks, fluid models, overload control, congestion collapse, switching dynamical systems, bi-stability, periodicity

activation and release thresholds (FQR-ART) overload control (specified in §2.1 below), which was developed in [28, 32]. The control is designed to automatically *switch* on sharing (serving some customers from the other pool) when an unexpected overload occurs, and switch off sharing when the overload incident is over, based only on the observed queue lengths. While the control is switched on, it aims to hold the two queues nearly fixed at some pre-specified ratio. From a many-server asymptotic perspective, the fixed-queue-ratio goal is to produce *state space collapse* (SSC).

It is significant that when the control parameters are chosen appropriately, the control is both effective and robust. In particular, it is successful in automatically switching on and off as needed, and in producing the desired SSC (again, automatically), *even under unrealistically-extreme conditions*; see [31] for key theory, involving an averaging principle, and §4.3 in [33] for important examples.

Nevertheless, in some extreme cases a performance degradation was demonstrated via simulation in [32]. More specifically, with *highly inefficient sharing* (the service rate is much less for the other customers), *if the control parameters are badly-chosen*, a system that is recovering from an overload incident, and is *no longer overloaded*, may get stuck in an oscillatory behavior that is due to unintended on-and-off switchings of the control. We now provide mathematical analysis that establishes key properties of this oscillatory behavior and provides a way to approximately quantify it. In [32] we developed a fluid approximation that can be used, in addition to simulation, to ensure that the bad behavior does not occur. Nevertheless, it is important to carefully study the limitations of controls. The insights gained should be useful for studying other overload controls.

A switching control. Most of the literature on control of queueing networks deals with ongoing operations, in which the control is operating continuously. Typically, it is also assumed that the arrival rates and total service capacity are known. However, here we are considering the control in [32] which automatically switches on and off, as was briefly described above. The fluid analysis we perform thus falls within the settings of (deterministic) *switching dynamical systems* [23].

A simple example of a switching system is the description of heated space. If the target temperature is set to be T_F°, then a thermostat should turn the heating on whenever the temperature drops below level T, and off when it reaches the target again. The *ideal* dynamic system's description then has two phases: The reaching (transient) phase, which describes the system until the temperature reaches level T, and the "sliding" phase, in which the temperature remains fixed ("slides") at T. Since heat is lost continuously, a

true sliding phase requires that the thermostat switch the heating on and off infinitely-many times during any time interval. In reality, a hysteresis control is employed, namely, the thermostat turns the heat on when the temperature reaches some level $T_\ell < T$, and off when the temperature hits some level $T_h > T$. Hence, a more realistic description of the corresponding dynamical system has the temperature chatter about level T, with this chattering being faster and smaller the more accurate the thermostat is. If the chattering is sufficiently small and fast, then the hysteresis dynamics approximate the ideal sliding phase quite well.

The queueing system we consider here has important similarities with the heating system describe above in that it is designed to "slide" on some region of its state space. More importantly, as we explain below, many other control mechanisms for queueing systems are designed for this purpose. The difference between the "ideal" sliding and the dynamics that are experienced in practice must be taken into account so as to avoid the harmful phenomena described here.

State space collapse, sliding motion and chattering. Asymptotic SSC in heavy-traffic limits is often a key step in developing effective (e.g., asymptotically optimal) controls for multidimensional stochastic networks; e.g., [4, 14, 15, 31, 34, 41, 43, 49]. (Related ideas date back to [46], but the systems there are uncontrolled.) As the term suggests, SSC means that the limit process is of a lower dimension than the prelimit process. More precisely, if SSC holds, then the limit process "collapses" (i.e., is confined) to a lower dimensional subset of its full state space. It is significant that SSC is often not only a mathematical tool that is employed to simplify asymptotic analysis, but rather, as in [31], SSC *may be a goal* of the control. We elaborate on the relation of SSC to optimal control in §9 below. See also page 136 in [1].

In the context of a *functional weak law of large numbers* (FWLLN) or *fluid limit*, asymptotic SSC corresponds to the limiting deterministic fluid process exhibiting a *sliding motion*, i.e., all the fluid trajectories "slide" on a lower-dimensional subspace, called a *sliding manifold*; see, e.g., §14.1 in [21] and §1.2.3 in [23]. In such cases, the fluid limit often has discontinuous dynamics in its full state space; i.e., it is governed by an *ordinary differential equation* (ODE) with a discontinuous right-hand side. The discontinuous dynamics is often avoided by assuming that the initial condition is asymptotically on the sliding manifold and restricting attention to the behavior of the limit on that region of the state space. However, if the initial condition of the fluid limit is not on the sliding manifold, the fluid trajectory must first go through a transient period before reaching the manifold; see Theorem 3 in [4] and

the explanation preceding it. (We remark that sliding manifolds should not be confused with the *invariant manifolds* in [4], which are defined to be the fixed points of the fluid limit. In particular, on the invariant manifold, the fluid trajectories are constant functions, whereas on a sliding manifold, the fluid limits may exhibit a transitory behavior.)

An effective SSC control must therefore (i) pull the system to the sliding manifold without undue delay and (ii) ensure that the system remains on the sliding manifold thereafter. For queueing networks, this may require specifying different routing rules for different regions of the state space - on and off the sliding manifold. For example, suppose that the state space $\mathbb{S}$ can be partitioned into three disjoint subsets $\mathcal{M}$, $\mathcal{M}^+$ and $\mathcal{M}^-$, where $\mathcal{M}$ is a sliding manifold, while $\mathcal{M}^+$ and $\mathcal{M}^-$ are "above" and "below" $\mathcal{M}$. A sliding-mode control will direct trajectories starting in $\mathcal{M}^-$ upwards toward $\mathcal{M}$, and downwards toward $\mathcal{M}$ from $\mathcal{M}^+$. Ideally, a sliding-mode control that starts in $\mathcal{M}^-$ will switch immediately once the fluid trajectory hits $\mathcal{M}$, aiming to keep that trajectory sliding on $\mathcal{M}$ after that hitting time. In reality, however, there may be a delay period until the control switches, so that the trajectory will cross immediately into $\mathcal{M}^+$ after hitting $\mathcal{M}$. Once the control finally switches, the trajectory is in $\mathcal{M}^+$ and the trajectory reverses its direction towards $\mathcal{M}$, but may again cross $\mathcal{M}$, this time into $\mathcal{M}^-$, because of delays in switching the control. This is the *chattering* phenomenon in the control literature; see §14.1 in [21]. When this chattering occurs, the sliding manifold $\mathcal{M}$ becomes a *switching manifold*, because the system switches its dynamics each time it crosses $\mathcal{M}$. Figure 1 depicts a schematic representation of chattering about a manifold $\mathcal{M}$, denoted by the dashed line, in the two-dimensional plane.

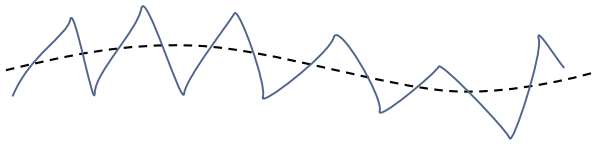

FIG 1. *Schematic depiction of Chattering (solid line) about a sliding manifold $\mathcal{M}$ (dotted line)*

The queueing context. Within queuing theory, the current paper should be considered in the context of instability of subcritical queueing networks, and in particular, instability caused by the control. Subcritical queueing networks that become overloaded due to exercising a bad control are said to experience *congestion collapse*, as in [39]; see §1 in [32] and §2.2 in [33].

The first studies of unstable subcritical queueing networks are the Lu-Kumar [25] and Rybko-Stolyar [37] networks. These networks were constructed as special, "atypical", examples with the purpose of demonstrating that sub-criticality is not sufficient for stability of queueing systems. It was only later acknowledged that these two counterexamples are indicative of general phenomena (see the discussion in §3 in [6]). For example, a two-station multiclass stochastic network, operating under FIFO, was shown to be unstable in [5]. The analysis there was carried out under the assumption that the number of classes is very large, namely, jobs move through station 2 a large number of times (e.g., 1,600 times) before exiting the system; see §2 in this reference. Nevertheless, a simulation experiment in [9] shows that with only four visits to station 2, the system is already unstable. Similarly, we consider extreme parameters for our instability analysis so as to permit qualitative and quantitative analysis of the fluid model, and use simulation to show that the oscillatory behavior can occur for the stochastic system with realistic parameters.

The setting. In this paper we illustrate the chattering phenomenon in a queueing network. Specifically, we consider a deterministic fluid approximation arising in the many-server heavy-traffic limit for a system with two service pools, each having its own arrival process and designated queue, that is operating under the FQR-ART overload control which we suggested in [32]. Normally, the two pools process work from their designated queues only. However, when an overload occurs due to an unexpected shift in the arrival rates, the control automatically identifies which queue should receive help and sharing begins, so that jobs from the overloaded queue are routed to both service pools, according to a routing rule that will be specified below.

The overload control was created for two call centers that normally operate separately, but might benefit by assisting each other to respond to unexpected overloads by temporarily serving some of the other customers. Given this call center motivation, we refer to pools of agents and the customers served in the other (not designated) pool as shared customers. When sharing is activated, the goal is to maintain the two queues nearly fixed at a pre-specified ratio during overload periods that is optimal with an appropriate cost formulation; see [28].

We showed that sharing can be effective even if sharing is inefficient, i.e., the shared customers are served at a slower rate. Since there is the possibility of performance degradation if there is too much sharing, it is necessary to choose the control parameters appropriately. The root cause of the chattering discussed here is indeed the combination of excessive inefficient sharing and

poorly chosen control parameters. To avoid excessive simultaneous sharing of customers in both directions ("two-way sharing,"see §4.1 in [28]), sharing with pool 1 helping queue 2 is activated only if the number of shared customers in pool 2 is below a certain (small) threshold, and similarly in the other direction. This latter restriction can cause delays in activating sharing when the direction of overload switches. Once activated, the control aims to produce asymptotic SSC by confining the queues to a certain region of the state space in the fluid limit [31]. In the fluid limit, this SSC translates to sliding motion on one of two sliding manifolds, each associated with one direction of sharing. We elaborate in §2 below.

When sharing is inefficient and the control parameters are not chosen appropriately, delays in activating the control can cause so much chattering that the fluid trajectory hits both sliding manifolds, without remaining in either, leading to complex chattering behavior. Here the chattering manifests itself in periodic oscillations, which lead to inefficient utilization of the service capacity. In turn, the inefficient utilization of agents creates severe overloads, even though *the arrival rates we consider are smaller than the potential service capacity.*

Chattering in sliding-mode controls is a well-known phenomenon in deterministic control theory. Indeed, chattering is considered to be the natural "state of affairs", whereas perfect sliding motion is considered ideal and unrealistic; e.g., §14.1 in [21]. Accordingly, even though we focus on a single system that operates under a specific control, our results have broader relevance. In particular, similar phenomena should be expected to occur with other SSC-inducing controls when there are deviations from ideal modeling assumptions, such as stationarity, or "convenient" initial conditions and control settings.

Switching dynamical systems. The chattering found in the fluid model implies that the ODE governing the evolution of the fluid trajectories switches whenever the control is activated or released. Therefore, the appropriate fluid model $x := \{x(t) : t \geq 0\}$ for the stochastic system is a *switching dynamical system* $\dot{x} = f_{\sigma(x)}(x)$, where $\sigma(x)$ achieves a finite set of values, f_i is a continuous function for each value i of σ, but the function f_σ is discontinuous [23]. As the notation suggests, the switching epochs are state dependent (depending only on the value of the solution x), so that the ODE is autonomous (time-homogeneous).

The framework of switching systems in general, and of systems with sliding motion in particular, is outside the classical ODE and dynamical-systems theory, because the right-hand side function f_σ is not continuous, and so it is not locally Lipshcitz. Hence, the conditions of the Picard-Lindelöf theorem,

ensuring the existence of a unique solution to the ODE, are not satisfied. In general, the existence of a unique solution to a switching system with no sliding motion can only hold in the Carathéodory sense, namely, such a solution is an absolutely-continuous function that satisfies the ODE almost everywhere; see [23]. A solution with a sliding motion is generally considered to hold in the Filippov sense [12], In [30, 31] we have proved that a unique solution exists for the fluid limit of our system during sliding motion via a stochastic *averaging principle* (i.e., the control achieves the desired asymptotic SSC).

Analytical contribution. In addition to exposing the chattering behavior discussed above, our current work has important analytical contributions. We emphasize at the outset that the derivation of the fluid model, which will also be shown to be the FWLLN in §I.2, and its quantitative analysis is relatively standard. The stronger analytical contributions lie in the nontrivial qualitative analysis of the fluid model. Specifically, we provide sufficient conditions for chattering to lead to endless oscillations, and prove the existence of a periodic equilibrium. Furthermore, we provide a simple algorithm to efficiently analyze the system for any given initial condition.

It is known that even seemingly simple switching systems can experience chaotic-like behavior, e.g., have infinitely-many periodic equilibria that are dense in the state space, and exhibit high sensitivity to perturbations of the initial condition (popularly known as "the butterfly effect"); see, e.g., [8, 11]. Such systems are clearly unamenable to long-run analysis. Even fluid models of *uncontrolled* systems can have uncountably-many periodic equilibria [24]. However, numerical experiments suggest that our system has at most one periodic equilibrium, and that it is bi-stable, i.e., any fluid trajectory can have long-run behavior of only two kinds: either it converges to the periodic equilibrium, or else it converges to the unique stationary point (which is therefore asymptotically stable).

To conduct a more complete study of the (bi)stability properties of the fluid model, we create an approximation to the fluid system. (Note that "stability" here does not refer to the prelimit queueing system which is always stable due to assumed abandonment.) For that approximating dynamical system we show that all oscillating solutions must converge to the unique periodic equilibrium (of the approximating system), while all other solutions converge to the unique stationary point, which is the same as that of the fluid limit. In particular, the approximating system is bistable. We conjecture that the same is true for the fluid limit (Conjecture 5.1 below), and support this conjecture by numerical experiments in §7.

To summarize, we develop and analyze two layers of approximations, one being the fluid limit, which approximates the stochastic system, and the other being an approximating dynamical system which serves as a simplified approximation to the fluid limit, whose qualitative behavior is easier to characterize.

Implications of the fluid analysis to the stochastic system. A straightforward implication of our result that the fluid limit may oscillate indefinitely is that the prelimit stochastic systems can experience congestion collapse. Moreover, the fluid limit may oscillate, even though the stochastic system in the pre-limit is an ergodic *continuous-time Markov chain* (CTMC) and is therefore necessarily aperiodic with a unique equilibrium (stationary) distribution. Since the CTMC converges to its unique stationary distribution also for initial conditions that are associated with oscillatory fluid limits, one concludes that the convergence rate of the CTMC to stationarity must be prohibitively slow. We elaborate in §8.

Our fluid analysis also has indirect implications to the stochastic system. Specifically, stochastic noise, which is not captured by the fluid approximation, may eventually push the system into the oscillatory behavior, even if the system is unambiguously initialized in the attraction region of the stationary point. This suggests that stochastic fluctuations can lead to *fluid-scaled fluctuations*. In addition, oscillations can occur in the stochastic system even if its fluid limit does not possess a periodic equilibrium, and never oscillates. Therefore, studying the relatively simple fluid model is important for gaining insight into the dynamics of the stochastic system. See the examples in §7.3 below.

Organization. The rest of the paper is organized as follows. We describe the stochastic model and the control in §2. In §2.2 we explain how to construct a direct fluid model to approximate the system's dynamics. The switching fluid model is derived in §3. Qualitative analysis, including relevant equilibrium and stability notions for dynamical systems, are rigorously defined and analyzed in §4. In §5 we show that the fluid model can oscillate indefinitely and when it does we show there exists a periodic equilibrium. The approximating dynamical system to the fluid model is developed in §6 and is shown to be bi-stable. Numerical examples and simulation experiments are provided in §7. In §8 we study the implications of the fluid analysis to the stochastic system, and in particular, to the long-run behavior of the underlying CTMC. General takeaways from our results, applicable to other systems and controls, are discussed in §9. We conclude in §10. Many of the results are proved in the appendix, and additional results appear in Sections F–J.

2. The model. We start by reviewing the stochastic model which is assumed Markovian, and in particular, it can be described as a CTMC. In §2.2 we quickly develop the deterministic fluid model to the stochastic system, which will be our focus in this paper. We defer the proof that the fluid model is indeed a rigorous approximation via a FWLLN to the appendix; see §I.2.

The model has two large service pools of many homogeneous agents in a call center, each with its own arrival stream and designated queue for waiting customers. We assume that customers have finite patience, and will abandon if their wait time in queue exceeds their patience. The two pools are designed to operate independently when both are normally loaded, i.e., to serve their own arrivals only, but all the agents can help both customer classes.

Sharing of customers (namely, routing customers from one pool to be served in the other pool) may be beneficial if one of the pools is overloaded, even if sharing makes the second service pool overloaded as well, because abandonment keep the two queues stable. Indeed, in [28] we showed that sharing of customers may be optimal during overload periods in a deterministic fluid approximation, assuming a convex holding cost is incurred on the two queues. However, as we showed in Proposition 2 in [28], when agents are less efficient in serving the other class, i.e., agents serve shared customers slower on average than their designated customers, it is never optimal to share in both directions simultaneously. Nevertheless, since sharing of customers in either direction takes place sometimes, and some sharing in both directions simultaneously may also take place, the routing graph of the system has the letter X shape, and is therefore called the X model in the call-center literature.

Figure 2 is a schematic portrayed of the X model; the circles represent the service pools, and the open-ended rectangle represent the buffers, and the arrows connecting the circles to the rectangles represent the allowed routing of customers to service. The solid arrows pointing to the buffers represent input (due to arrivals), and the dotted arrows represent output (due to abandonment from the buffers, and service completions from the service pools). We also show a figure of the N model in Figure 3 in which sharing of customers is possible in one direction only. We discuss related known results concerning the N model in §9 below. (Figure 3 has no arrows from the buffers since the N systems we discuss have no abandonment.)

In general, there is a fluid-optimal amount of sharing for any given pair of arrival rates and so, to find how many agents in the helping pool should be assigned to shared customers requires knowing the exact arrival rates

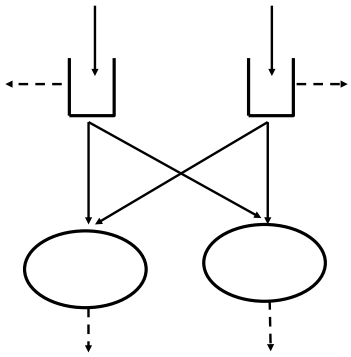

FIG 2. *The X model*

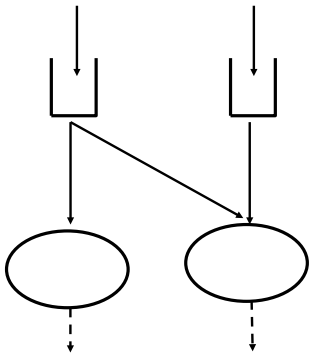

FIG 3. *The N model*

during the overload period. A simplification is achieved by observing that the exact amount of sharing does not need to be determined at the outset, since it can be achieved, at least approximately, if the two fluid queues are kept at a fixed ratio during overload periods. We again refer to [28]. There is a different optimal ratio for each direction of sharing, and the direction of sharing depends on which pool is overloaded.

The above reasoning lead us to design the *fixed queue ratio with thresholds* (FQR-T) overload control, which (i) is activated automatically once the queue ratio exceeds a certain "activation threshold" (so that the system is considered overloaded); (ii) aims to maintain the two queues at a pre-specified fixed ratio (in the many-server asymptotics); (iii) class-i customers are routed to pool j only if there are no class j customers in pool i, $i \neq j$.

In time-varying settings, the direction of overload may switch, so that the direction of sharing must switch as well. If the *one-way sharing rule* in Condition (iii) above is forced, then substantial delays in switching the direction of sharing may occur. We therefore modified FQR-T in [32] by introducing *release thresholds* for the service process. Specifically, in the modified *fixed queue ratio with activation and release thresholds* (FQR-ART) control the one-way sharing rule is relaxed as follows: class-1 customers can be routed to pool 2, provided that the number of class-2 customers in pool 1 is smaller than a release threshold $\tau_{2,1} > 0$, and similarly in the other direction. We elaborate in §2.1 below.

Cyclic routing graph. An important characteristic of the X model is that its (undirected) routing graph is cyclic. In particular, it is the most basic cyclic *parallel server system* (PSS). The X model is therefore easier to study than other cyclic PSS's but at the same time serves as a representative to problems that are associated with its cyclic structure. Indeed, in [28] we

showed that the QIR control from [15] can produce severe congestion collapse if applied to the X model when the service rates of shared customers are slower than those of designated ones. This congestion collapse cannot occur in PSS's having a tree graph; see Theorem 3.1 in [15].

As was mentioned above, the FQR-ART control aims to avoid simultaneous sharing of customers as much as possible, and to reduce the system into an N model (as in Figure 3) at any given time that sharing take place, although the some simultaneous sharing is possible, and the direction of sharing may change with time. It is therefore compelling to compare our results regarding the X model to known results on the well-studied N model; see, e.g, [2, 18, 20, 45] and references therein. In §9 we make such comparisons to indicate how our results here, as well as results from our previous work on the X model, provide important insights to other SSC-inducing controls, taking the N model as an example. We note that the N model has the most basic tree structure of a PSS with more than one class of arrivals and more than one service station, making it a "representative model" for PSS's with tree structures (in a similar manner to the X model being a "representative model" for cyclic PSS's). In particular, our insights are not restricted to PSS's with cyclic routing graphs.

2.1. *The FQR-ART control.* We will start by developing a deterministic fluid approximation for the stochastic system directly, but to fully describe the control, we must consider that fluid model from an asymptotic perspective. We therefore consider a sequence of X models indexed by superscript n, where system n has m_i^n agents in pool i and arrival rate λ_i^n of class-i customers, $i = 1, 2$. We assume that the arrival rates and number of agents in each pool grow proportionally to n as $n \to \infty$, putting us in the many-server heavy-traffic framework.

The control of each system $n \geq 1$ is based on two *activation thresholds*, $k_{1,2}^n$ and $k_{2,1}^n$, two *release thresholds*, $\tau_{1,2}^n$ and $\tau_{2,1}^n$, and two ratio parameters $r_{1,2}$ and $r_{2,1}$. These ratios, which are independent of n, are chosen to be optimal in a fluid model of an overloaded system (here we will consider underloaded systems), as was mentioned above.

Let $Q_i^n(t)$ denote the number of class-i customers waiting in their designated queue at time t, and let $Z_{i,j}^n(t)$ denote the number of class-i customers being served in pool j at time t. The FQR-ART is an overload control, namely, it is designed to be activated and start customer sharing automatically when an overload occurs. To define overloads, we consider the difference processes. For $t \geq 0$,

$$D_{1,2}^n(t) \equiv Q_1^n(t) - r_{1,2}Q_2^n(t) - k_{1,2}^n, \quad D_{2,1}^n(t) \equiv r_{2,1}Q_2^n(t) - Q_1^n(t) - k_{2,1}^n. \tag{2.1}$$

As long as $D^n_{1,2} < 0$ and $D^n_{2,1} < 0$, the system is considered normally loaded. Once one of these difference processes hits 0, which corresponds to the ratio between the two queues hitting one of the activation thresholds, the system is deemed overloaded, and sharing begins, provided that there is only a small number of shared customers in the overloaded pool. By "small number" we mean that the number of shared customers in the overloaded pool is no larger than its associated release threshold. For example, if $D^n_{1,2}(t) \geq 0$, then class 1 is judged to be overloaded (because then $Q^n_1(t) - r_{1,2}Q^n_2(t) \geq k^n_{1,2}$) and it is desirable to send class-1 customers to be served in pool 2. However, sharing is allowed only if $Z^n_{2,1}(t) \leq \tau^n_{2,1}$. Similar rules apply to overloads in the other direction. (It is important that $\tau^n_{i,j}$ are taken to be small numbers, so that not too much harmful simultaneous sharing can occur. However, these threshold must be strictly positive; see (3.21) below and §3 in [32].)

Once sharing is activated, say with class 1 receiving help from pool 2, the routing rule is as follows: Any agent, from either pool, that becomes available at any time t, will take his next customer from class 1 if $D^n_{1,2}(t) > 0$, and will take his next customer from his designated queue otherwise. Observe that this means that agents from pool 1 will only take customers from their own queue, but some class 1 customers will be routed to pool 2. The routing mechanism when class 2 is overloaded is similar, with $D^n_{2,1}$ replacing $D^n_{1,2}$, and the labels of the thresholds switched.

2.2. *A deterministic fluid model.* If the arrival processes are independent Poisson processes, and all service times and times to abandon are independent exponential random variables, then the six-dimensional process

$$X^n(t) = (Q^n_i(t), Z^n_{i,j}(t); i, j = 1, 2), \quad t \geq 0, \tag{2.2}$$

is a CTMC. Our goal is to develop and then analyze a fluid approximation for this CTMC, based on asymptotic considerations (which will be made rigorous in §I.2).

When sharing is active, the control aims to keep the two queues at the corresponding fluid-optimal ratio, either $r_{1,2}$ or $r_{2,1}$, depending on the direction of sharing. Minor modifications to the statement and proof of Corollary 4.1 in [31] show that, if the system is overloaded and there is no sharing initially, then the control achieves asymptotic SSC in the fluid limit (or under any scaling of the appropriate process in (2.1) that is larger than $\log n$). More general assumptions were considered in [32]. The mathematical support for the asymptotic SSC was a direct consequence of the aforementioned stochastic averaging principle.

The oscillatory performance and its resulting congestion collapse we analyze here does not involve the averaging principle, because there is no SSC.

Indeed, unlike the fluid models in [31] and [32], the fluid model we develop here has an explicit solution. The challenges are associated with proving that oscillations (and congested collapse) can be self-sustained and in studying the long-run behavior of the fluid model.

It is significant that the fluid approximation for X^n is obtained as the FWLLN for $\bar{X}^n \equiv X^n/n$, see §I.2. However, we start by deriving the fluid model directly. (We refer to the fluid model as fluid approximation or limit, depending on the context, as the terms are equivalent in our case.) For each of the six stochastic processes comprising X^n in (2.2) there is a fluid counterpart, namely a deterministic and almost-everywhere differentiable function. We let $x \equiv \{x(t) : t \geq 0\}$ denote the fluid approximation of X^n, where

$$x(t) = (q_1(t), q_2(t), z_{1,1}(t), z_{1,2}(t), z_{2,1}(t), z_{2,2}(t)), \quad t \geq 0,$$

and call a time t "regular" if $x(t)$ is differentiable at t. In our case, any compact interval will have at most a finite number of points that are not regular.

To derive the fluid equations, we simply replace the instantaneous rates of the stochastic processes at each time t with instantaneous rates of change of the derivatives of their fluid counterparts, e.g., the instantaneous rate of abandonment from queue 1 at time t in system n is $\theta_1 Q_1^n(t)$, which becomes $\theta_1 q_1(t)$ in the fluid model. Similarly, the instantaneous rate of departure from service in pool j at time t is $\mu_{j,j} Z_{j,j}^n(t) + \mu_{i,j} Z_{i,j}^n(t)$ in system n is replaced with the instantaneous processing rate $\mu_{j,j} z_{j,j}(t) + \mu_{i,j} z_{i,j}(t)$ in the fluid model. Combining all these instantaneous rates gives the derivative of $x(t)$ at a regular time t.

For example, if both queues are smaller than the activation thresholds at a time t, then any newly-available agent in pool 1 will take his next customer from queue 1 in the stochastic system. Similar reasonings applied to q_2 give that, if $q_1(t) < k_{1,2}$ and $q_2(t) < k_{2,1}$, and t is regular, then

$$\begin{aligned} \dot{q}_1(t) &= \lambda_1 - \theta_1 q_1(t) - \mu_{1,1} z_{1,1}(t) - \mu_{2,1} z_{2,1}(t), \\ \dot{q}_2(t) &= \lambda_2 - \theta_2 q_2(t) - \mu_{2,2} z_{2,2}(t) - \mu_{1,2} z_{1,2}(t). \end{aligned} \tag{2.3}$$

We derive the full set of differential equations for the fluid model during overload periods (due to congestion collapse) in §3.1 below.

The purpose of FQR-ART is to produce SSC in the fluid limit by sending customers from one queue to both pools according to the routing rules described above during overload periods. If the control is successful in achieving SSC, the six-dimensional fluid model is confined to one of the *sliding*

manifolds

$$\mathbb{S}_{1,2} \equiv \{x \in \mathbb{S} : q_1 - r_{1,2}q_2 = k_{1,2},\ z_{1,1} + z_{2,1} = m_1,\ z_{1,2} + z_{2,2} = m_2\},$$
$$\mathbb{S}_{2,1} \equiv \{x \in \mathbb{S} : r_{2,1}q_2 - q_1 = k_{2,1},\ z_{1,1} + z_{2,1} = m_1,\ z_{1,2} + z_{2,2} = m_2\},$$

where $\mathbb{S} = \mathbb{R}^2_+ \times [0, m_1] \times [0, m_2]$ is the domain of x.

We note that with each of the two sliding manifolds in (2.2) there is an associated fixed point to which the "sliding" fluid solutions converge as $t \to \infty$. In particular, letting $x^*_{1,2}$ denote the stationary point on $\mathbb{S}_{1,2}$ and assuming that $x(0) \in \mathbb{S}_{1,2}$ during an overload period with sliding motion on $\mathbb{S}_{1,2}$, we have shown in [30] that $x(t) \to x^*_{1,2}$ as $t \to \infty$, i.e., $x^*_{1,2}$ is globally asymptotically stable. If $x(0) = x^*_{1,2}$, then $x(t) = x^*_{1,2}$ for all t, so that the set $\{x^*_{1,2}\}$ is the invariant manifold for the fluid model (sliding on $\mathbb{S}_{1,2}$), as in [4].

The behavior of the fluid limit when sliding on one of these manifolds can be thought of as an infinitely-fast chattering with infinitely-small fluctuations of the queues about the corresponding activation threshold. This view can be justified rigorously via the aforementioned stochastic averaging principle; see §4 in [30] and Theorem 4.1 in [31].

Observe that the fluid model is essentially a *three-dimensional process* on either one of these sliding manifolds, because knowing $x_3 \equiv (q_1, z_{1,2}, z_{2,1})$ for example, is sufficient to determine the value of the remaining three processes. Here, however, we are interested in bad oscillatory behavior when the fluid model overshoots past the sliding manifold due to delay in activating the control, where a delay is caused if $z_{j,i}(t_0) > \tau_{j,i}$, at the time t_0 in which $\mathbb{S}_{i,j}$ is hit. If no SSC occurs, we must consider all six components of the fluid model and, as will become clear below, four different switching epochs for each cycle. We can obtain considerable simplification by considering a symmetric model. Symmetry reduces the amount of notation and, as will become clear later, allows us to focus attention on two switching times in each cycle instead of four.

A symmetric model. In order to expose the bad behavior that can result from poorly chosen controls, we consider a special case that is easier to analyze than the general model. In particular, we consider systems with the following parameters

(2.4)
$$\mu_{1,1} = \mu_{2,2} = 1, \quad \mu_{1,2} = \mu_{2,1} = \mu < 1, \quad \lambda_1 = \lambda_2 = \lambda < 1, \quad \theta_1 = \theta_2 = \theta > 0,$$
$$m_1 = m_2 = 1, \quad r_{1,2} = r_{2,1} = 1, \quad \tau_{1,2} = \tau_{2,1} = \tau > 0, \quad k_{1,2} = k_{2,1} = \kappa.$$

Observe that time is measured in terms of $\mu_{1,1}$ and $\mu_{2,2}$ (which are normalized to be equal to 1). In this model there are 5 parameters instead of 16

in the general case. There is the triple of model parameters (λ, μ, θ) and the pair of control parameters (κ, τ). Note that each of the pools is underloaded if there is no sharing that slows its potential service capacity, because $\lambda_i < \mu_{i,i} m_i = 1$, $i = 1, 2$.

In this model, there is sharing with all class-2 fluid sent to pool 1 if $q_2(t) > q_1(t) + \kappa$ and $z_{1,2}(t) \leq \tau$; there is sharing with all class-1 fluid sent to pool 2 if $q_1(t) > q_2(t) + \kappa$ and $z_{2,1}(t) \leq \tau$; there is complex sharing, associated with sliding motion and described by the averaging principle if $q_2(t) = q_1(t) + \kappa$ and $z_{1,2}(t) \leq \tau$ or if $q_1(t) = q_2(t) + \kappa$ and $z_{2,1}(t) \leq \tau$; there is possibly sharing according to the spare capacity control described above if $q_1(t) \geq \kappa$ and $z_{1,2}(t) + z_{2,2}(t) < m$ or $q_2(t) \geq \kappa$ and $z_{2,1}(t) + z_{1,1}(t) < m$. otherwise there is no sharing actively taking place.

We have assumed in (2.4) that $\lambda < 1$, so that either pool is underloaded if it serves its own class only (because $\mu_{i,i} m_i = 1$, $i = 1, 2$). It will be convenient to assume that $\lambda \leq 1 - \tau$. In that case, if class i fluid is sent to pool j at time t, $i \neq j$, then $z_{j,i}(t) \leq \tau$ and the instantaneous service rate in pool i is

$$\mu z_{j,i}(t) + z_{i,i}(t) = \mu z_{j,i}(t) + (1 - z_{j,i}(t)) \geq \mu z_{i,j}(t) + 1 - \tau \geq \mu z_{i,j}(t) + \lambda \geq \lambda,$$

implying that the instantaneous total service rate in pool i is larger than the arrival rate to that pool so that q_i is decreasing; see also (2.3). In addition, to achieve explicit solutions to the ODE's we develop, we will assume that $\theta < \mu$. We summarize in the following assumption.

ASSUMPTION 1. *The model parameters satisfy* (2.4). *Furthermore,* $\lambda \leq 1 - \tau$ *and* $\theta < \mu$.

Assumption 1 is not necessary for chattering and oscillations to occur, and is taken in order to somewhat simplify the analysis. Since τ is small, the condition $\lambda \leq 1 - \tau$ is a slight strengthening of the condition $\lambda < 1$ in (2.4), and implies that either pool is underloaded when its own class of customers receives help from the second pool, *regardless of the value of* μ. To see this, observe that the instantaneous total service rate at pool j at time t, if there is no routing of new class-i customers to that pool, is $\mu z_{i,j}(t) + (1 - z_{i,j}(t))$, and that

$$\mu z_{i,j}(t) + (1 - z_{i,j}(t)) \geq \mu z_{i,j}(t) + 1 - \tau \geq \mu z_{i,j}(t) + \lambda > \lambda.$$

The condition $\theta < \mu$ simplifies the exposition of the fluid model. Specifically, As will be seen below, we provide closed-form solutions for the fluid model which are not defined for $\theta = \mu$, e.g., see (3.6). Therefore, one needs to solve

for the case $\theta = \mu$ separately. Letting $x(\cdot, \theta, \mu) := \{x(t; \theta, \mu) : t \geq 0\}$ denote our solution for the fluid equations as a function of θ and μ, and defining $x(t, \theta, \theta)$ and $x(t, \mu, \mu)$ to be the solution for the fluid equations when $\theta = \mu$, it easily checked that our explicit solution $x(\cdot, \theta, \mu)$ is continuous in θ and in μ. We further remark that our solution remains valid if $\theta > \mu$, but the sign of some arguments changes.

Since the activation thresholds κ are strictly positive in the fluid model, there is no ambiguity about the translation of the FQR-ART control to the fluid model when there is no SSC. It is then entirely determined by the processes

$$d_{1,2}(t) = q_1(t) - q_2(t) - \kappa \quad \text{and} \quad d_{2,1}(t) = q_2(t) - q_1(t) - \kappa, \quad t \geq 0, \tag{2.5}$$

which are simply the fluid counterparts of (2.1). Due to the assumed symmetry, the state space of the fluid model is $\mathbb{R}_+^2 \times [0,1]^4$ and the sliding manifold are defined via

$$\begin{aligned} \mathbb{S}_{1,2} &\equiv \{x \in \mathbb{S} : d_{1,2} = 0,\ z_{1,1} + z_{2,1} = 1,\ z_{1,2} + z_{2,2} = 1\} \\ \mathbb{S}_{2,1} &\equiv \{x \in \mathbb{S} : d_{2,1} = 0,\ z_{1,1} + z_{2,1} = 1,\ z_{1,2} + z_{2,2} = 1\}. \end{aligned} \tag{2.6}$$

For $i, j = 1, 2$, $i \neq j$, we define $\mathbb{S}_{i,j}^- \equiv \{x \in \mathbb{S} : d_{i,j} < 0\}$ and $\mathbb{S}_{i,j}^+ \equiv \{x \in \mathbb{S} : d_{i,j} > 0\}$.

If $x(t) \in \mathbb{S}_{i,j}$ for all t over some interval I, then x is said to slide on the sliding manifold $\mathbb{S}_{i,j}$. Chattering corresponds to the fluid trajectory hitting and immediately crossing a sliding manifold, e.g., when it is moving from $\mathbb{S}_{i,j}^-$ to $\mathbb{S}_{i,j}^+$ (necessarily via $\mathbb{S}_{i,j}$) without sliding on $\mathbb{S}_{i,j}$, and back from $\mathbb{S}_{i,j}^+$ to $\mathbb{S}_{i,j}^-$. It will be clear that chattering about one sliding manifold is not sustainable unless the fluid trajectory makes it all the way to the second manifold. When both manifolds are hit, we say that the fluid oscillates. Since we will consider initial conditions in $\mathbb{S}_{2,1}^+$, a full cycle is considered to end when the fluid trajectory first enters $\mathbb{S}_{2,1}^+$ after hitting $\mathbb{S}_{1,2}$. When chattering or oscillations occur, the sliding manifolds in (2.6) become *switching surfaces*, because the dynamics of the fluid model switches when it hits either of these subspaces.

The state space. It is easily seen from (2.3) that $\dot{q}_i(t) \leq \lambda - \theta q_i(t)$, and that this inequality holds for all $t \geq 0$ regardless of the routing. It follows from the comparison principle for ODE's, e.g., Lemma 3.4 in [21], that for all $t > 0$, $q_i(t) \leq \max\{q_i(0), \lambda/\theta\}$, $i = 1, 2$, and that, if $q_i(0) > \lambda/\theta$, then q_i must be strictly decreasing as long as $q_i(t) > \lambda/\theta$. Furthermore, q_i can never cross λ/θ from below, i.e., if $q_i(s) < \lambda/\theta$, then $q_i(t) < \lambda/\theta$ for all $t > s \geq 0$. We can therefore assume without any loss of generality that $q_i(0) < \lambda/\theta$

so that the state space of the symmetric model is the compact and convex subset $\mathbb{S} \subset \mathbb{R}_6$, where

$$\mathbb{S} \equiv [0, \lambda/\theta]^2 \times [0,1]^4. \tag{2.7}$$

3. The switching fluid model. Consider a system that has just recovered from an overload, in which class 1 was receiving help from pool 2. Suppose that λ_1, which was greater than $\mu_{1,1} m_1 = 1$ during the preceding overload period, dropped to the value $\lambda < 1$ in (2.4) Since sharing was taking place with pool 2 helping, we necessarily had $z_{2,1} < \tau$ and $q_1 - q_2 = \kappa > 0$ (x sliding on $\mathbb{S}_{1,2}$) during the overload period.

Assuming that $z_{1,2}$ was larger than τ during the preceding overload period, we designate by 0 the first time that $z_{1,2}$ hits τ, so that sharing can begin with pool 1 helping queue 2 if $d_{2,1}(0) > \kappa$. Formally,

Assumption 2 (initial condition).

$$x(0) \in \mathbb{S}, \quad q_1(0) > 0, \quad d_{2,1}(0) > 0, \quad z_{1,2}(0) = \tau \quad \textit{and} \quad 0 \le z_{2,1}(0) < \tau.$$

To describe the oscillatory behavior of the fluid model, we define the times

$$\begin{aligned} &T_1 \equiv \inf\{t \ge 0 : d_{2,1}(t) \le 0\}, \quad T_2 \equiv \inf\{t \ge 0 : z_{2,1}(\Sigma_1 + t) \le \tau\}, \\ &T_3 \equiv \inf\{t \ge 0 : d_{1,2}(\Sigma_2 + t) \le 0\}, \; T_4 \equiv \inf\{t \ge 0 : z_{1,2}(\Sigma_3 + t) \le \tau\}, \end{aligned} \tag{3.1}$$

where,

$$T_0 \equiv \Sigma_0 \equiv 0, \quad \Sigma_k \equiv \sum_{i=0}^{k} T_i \quad \text{and} \quad \mathcal{I}_i \equiv [\Sigma_{i-1}, \Sigma_i), \quad k = 1,2,3,4. \tag{3.2}$$

Observe that T_1 and T_3 are the hitting times of the switching manifolds, and are in fact the crossing times from above of these manifolds. At those hitting times, ongoing sharing ends. The times T_2 and T_4 are the hitting times (again, crossing times from above) of the release thresholds, so that sharing can begin at those times. For example, if $x(T_2) \in \mathbb{S}^+_{1,2}$, then sharing of class-1 fluid can begin at this time. See §3.2 below.

We refer to the times Σ_i as *switching times*, and to T_i as *holding times* (the times between switching). The length of each interval $\mathcal{I}_i$ is T_i, i.e., $|\mathcal{I}_i| \equiv \Sigma_i - \Sigma_{i-1} = T_i$, $1 \le i \le 4$. We will interchangeably write T_1 or Σ_1, and $T_1 + T_2$ or Σ_2, as convenient.

Clearly $T_1 > 0$ for the initial condition in Assumption 2, but it is possible that $T_i = 0$ for $i > 1$. Observe that if at the end of the first cycle $x(\Sigma_4)$

satisfies the same conditions specified for $x(0)$ in Assumption 2, then $x(\Sigma_4)$ can be taken as a new "initial condition" for the fluid model (which is time homogeneous, as will be shown below), and a new cycle begins. Furthermore, if both fluid queues are strictly positive on $[0, \Sigma_q)$ and $z_{2,1}(\Sigma_1) > \tau$ in addition to $d_{1,2}(\Sigma_2) > 0$, then $x(\Sigma_2)$ can be thought of as a "mirror image" of $x(0)$ because we necessarily have $0 < z_{1,2}(\Sigma_2) < \tau$. In particular $x(\Sigma_2)$ satisfies the conditions in Assumption 2, *but with the labels (subscripts) reversed*. Similarly, if both queues remain positive throughout $[0, \Sigma_3)$, then $x(\Sigma_3)$ is a "mirror image" of $x(\Sigma_1) \equiv x(T_1)$. This observation greatly simplifies the search for a periodic equilibrium since, on the trajectory of a periodic equilibrium, it holds that $x_s(\Sigma_2) = x(0)$ and $x_s(\Sigma_3) = x(\Sigma_1)$, where $x_s := (q_2, q_1, z_{2,2}, z_{2,1}, z_{1,2}, z_{1,1})$ (i.e., x_s has the labels of x reversed). We can then focus on analyzing a half cycle $[0, \Sigma_2]$ for the symmetric model.

Hence, we consider the fluid model as long as the conditions in Assumption 2 hold in the switching times, either for x or for x_s. It will be seen below that, for any initial condition in $\mathbb{S}$, $0 \le z_{i,j} \le 1$, $i, j = 1, 2$. However, the equations for q_1 and q_2 can become negative. We thus consider the fluid model on $[0, \Sigma_q)$, where

$$\Sigma_q \equiv \inf\{t > 0 : \min\{q_1(t), q_2(t)\} = 0\}. \tag{3.3}$$

Since $T_1 > 0$ for any initial condition satisfying Assumption 2, we necessarily have $\Sigma_1 > T_1 > 0$. Similarly, if $\Sigma_2 > 0$, then necessarily $T_3 > 0$. It follows that, if $\Sigma_q < \Sigma_4$, then $\Sigma_q \in \mathcal{I}_2$ or $\Sigma_q \in \mathcal{I}_4$. On the other hand, if $x(\Sigma_4)$ satisfies the conditions in Assumption 2, then $\Sigma_q > \Sigma_4$. We then take $x(\Sigma_4)$ as the initial condition for the second cycle, and start over. We will provide sufficient conditions for Σ_q to be infinite, in which case cycle-end time Σ_4 is the beginning of a new full cycle, and the fluid model keeps oscillating indefinitely. Since both queues are strictly positive throughout (despite Assumption 1), we get congestion collapse that is due to self-sustained oscillations.

3.1. *The switching fluid equations.*

3.1.1. *The equations on $\mathcal{I}_1$: Both pools serve queue 2 only.* Recall that over the interval $\mathcal{I}_1 \equiv [0, \Sigma_1)$ sharing takes place with both pools accepting only fluid from queue 2 and no fluid from queue 1. For a given initial condition $x(0)$ satisfying Assumption 2, and determined by specifying the triple $(q_1(0), q_2(0), z_{2,1}(0))$, the fluid equations for the service process are therefore

$$\dot{z}_{1,1}(t) = -z_{1,1}(t)\mu_{1,1}, \quad z_{2,1}(t) = 1 - z_{1,1}(t) \quad \text{and} \quad \dot{z}_{1,2}(t) = -z_{1,2}(t)\mu_{1,2},$$

so that

$$z_{1,1}(t) = (1 - z_{2,1}(0))e^{-t}, \quad z_{2,1}(t) = 1 - (1 - z_{2,1}(0))e^{-t}, \tag{3.4}$$
$$z_{1,2}(t) = \tau e^{-\mu t} \quad \text{and} \quad z_{2,2}(t) = 1 - \tau e^{-\mu t},$$

and the fluid equations for the queue processes are

$$\begin{aligned} \dot{q}_1(t) &= \lambda - q_1(t)\theta, \\ \dot{q}_2(t) &= \lambda - q_2(t)\theta - z_{1,1}(t)\mu_{1,1} - z_{2,1}(t)\mu_{2,1} - z_{1,2}(t)\mu_{1,2} - z_{2,2}(t)\mu_{2,2} \\ &= \lambda - q_2(t)\theta - [(1 - z_{2,1}(0))e^{-t} + 1 - \tau e^{-\mu t}] \\ &\quad - [1 - (1 - z_{2,1}(0))e^{-t} + \tau e^{-\mu t}]\mu \\ &= (\lambda - 1 - \mu) - q_2(t)\theta - (1 - \mu)(1 - z_{2,1}(0))e^{-t} + (1 - \mu)\tau e^{-\mu t}. \end{aligned} \tag{3.5}$$

For the given initial condition $x(0)$, we can calculate the interval termination time T_1 and the fluid performance functions in $\mathcal{I}_1$. Observe that by first solving for the service processes in (3.4), the autonomous (time-homogeneous) ODE for the queues becomes a nonhomogeneous first-order linear ODE. Under the condition $\theta < \mu$ in Assumption 1, the explicit solution to the ODEs (3.5) over $[0, T_1)$ is

$$\begin{aligned} q_1(t) &= q_1(0)e^{-\theta t} + \left(\frac{\lambda}{\theta}\right)(1 - e^{-\theta t}) \\ q_2(t) &= q_2(0)e^{-\theta t} + \left(\frac{\lambda - 1 - \mu}{\theta}\right)(1 - e^{-\theta t}) \\ &\quad - \left(\frac{(1 - \mu)(1 - z_{2,1}(0))}{1 - \theta}\right)(e^{-\theta t} - e^{-t}) + \left(\frac{(1 - \mu)\tau}{\mu - \theta}\right)(e^{-\theta t} - e^{-\mu t}). \end{aligned} \tag{3.6}$$

We see that $q_1(t)$ is strictly increasing in $\mathbb{S}$ and necessarily remains strictly positive in the interval $\mathcal{I}_1$. Given the initial conditions in Assumption 2 and the definition of $\Sigma_1 \equiv T_1$ in (3.1), this implies that both fluid queue lengths are necessarily strictly positive in the interval $\mathcal{I}_1$, so that $\Sigma_q > T_1$.

3.1.2. *The equations on $\mathcal{I}_2$: No active sharing.* Given any initial condition $(q_1(0), q_2(0), z_{2,1}(0))$, we can calculate T_1 and the 6-tuple $(q_i(T_1), z_{i,j}(T_1)); i, j = 1, 2)$. These provide the initial condition for the second interval $\mathcal{I}_2 \equiv [\Sigma_1, \Sigma_2)$. We assume that $z_{2,1}(T_1) > \tau$ so that sharing with pool 2 helping queue 1 did not begin at time T_1 and so $T_2 > 0$. The fluid equations

for the service process for $t \in \mathcal{I}_2$ are

(3.7)
$$\dot{z}_{2,1}(t) = -z_{2,1}(t)\mu_{2,1}, \quad \text{so that} \quad z_{2,1}(T_1+t) = [1-(1-z_{2,1}(0))e^{-T_1}]e^{-\mu t}$$
$$\text{and} \quad z_{1,1}(T_1+t) = 1 - z_{2,1}(T_1+t) = 1 - [1-(1-z_{2,1}(0))e^{-T_1}]e^{-\mu t}$$
$$\dot{z}_{1,2}(t) = -z_{1,2}(t)\mu_{1,2}, \quad \text{so that} \quad z_{1,2}(T_1+t) = \tau e^{-\mu(T_1+t)}$$
$$\text{and} \quad z_{2,2}(T_1+t) = 1 - z_{1,2}(T_1+t) = 1 - \tau e^{-\mu(T_1+t)}.$$

As long as both queues remain positive, since there is no new sharing in this second interval $\mathcal{I}_2$, at time T_1+t for $t \in [0, T_2]$, the queues evolve as follows:

$$\begin{aligned} \dot{q}_1(T_1+t) &= \lambda - q_1(T_1+t)\theta - z_{1,1}(T_1+t)\mu_{1,1} - z_{2,1}(T_1+t)\mu_{1,2} \\ &= -(1-\lambda) - q_1(T_1+t)\theta + (1-\mu)z_{2,1}(T_1)e^{-\mu t} \\ \dot{q}_2(T_1+t) &= \lambda - q_2(T_1+t)\theta - z_{2,2}(T_1+t)\mu_{2,2} - z_{2,1}(T_1+t)\mu_{2,1} \\ &= -(1-\lambda) - q_2(T_1+t)\theta + (1-\mu)z_{1,2}(T_1)e^{-\mu t} \end{aligned} \tag{3.8}$$

under the new initial condition $(q_1(T_1), q_2(T_1), z_{1,2}(T_1), z_{2,1}(T_1))$.

Paralleling (3.6), we can solve these ODE's explicitly: For all $t \in [0, T_2)$

(3.9)
$$q_1(T_1+t) = q_1(T_1)e^{-\theta t} + \frac{\lambda-1}{\theta}(1-e^{-\theta t}) + \frac{(1-\mu)z_{2,1}(T_1)}{\mu-\theta}(e^{-\theta t} - e^{-\mu t})$$
$$q_2(T_1+t) = q_2(T_1)e^{-\theta t} + \frac{\lambda-1}{\theta}(1-e^{-\theta t}) + \frac{(1-\mu)z_{1,2}(T_1)}{\mu-\theta}(e^{-\theta t} - e^{-\mu t}),$$

provided that $T_1 + t \le \Sigma_q$.

3.1.3. *The switching fluid model.* The equations on $\mathcal{I}_3 \equiv [\Sigma_2, \Sigma_3)$ and $\mathcal{I}_4 \equiv [\Sigma_3, \Sigma_4)$ are derived similarly to the equations for the intervals $\mathcal{I}_1$ and $\mathcal{I}_2$, assuming $\Sigma_q < \Sigma_4$. We summarize in the following definition of the direct fluid model. As was mentioned before, we consider the interval $[0, \Sigma_q)$ and provide sufficient conditions for Σ_q to be infinite. We further prove that oscillations must end at time Σ_q when this time is finite.

For two real numbers a, b, let $a \wedge b \equiv \min\{a, b\}$. We will later also use the notation $a \vee b$ for the maximum between the two numbers.

DEFINITION 3.1 (switching symmetric fluid model). *For any initial condition $x(0)$ satisfying Assumption 2, the fluid model for the symmetric system is the solution $x \equiv \{x(t) : t \in [0, \Sigma_4 \wedge \Sigma_q)\}$ to the autonomous (time invariant) switching ODE*

$$\dot{x} = f_{\sigma(x)}(x), \quad \sigma(x(t)) = 1, 2, 3, 4; \tag{3.10}$$

where f_1 is defined in (3.4)–(3.5), f_2 is defined in (3.7)–(3.8), f_3 satisfies the equations of f_1, but with the labels of the processes reversed, and f_4 satisfies and equations of f_2, with the labels of the processes reversed. The switching times Σ_i, $1 \le i \le 4$, *are determined by the value of the solution $x(t)$ at time t and are defined in (3.2). Furthermore, all points $t \in [0, \Sigma_4 \wedge \Sigma_q)$, except for the switching times, are regular.*

We refer to any specific solution to (3.10) as a fluid solution or a trajectory. As was mentioned above, if $x(\Sigma_4)$ satisfies Assumption 2, then it serves as an initial condition for the following cycle, so that (3.10) describes the fluid dynamics beyond the first cycle in an obvious way. In §I.2 we will show that the unique solution x to (3.10) with a given initial condition arises as the FWLLN of $\bar{X}^n$ in (2.2) as $n \to \infty$ over any compact subinterval of $[0, \Sigma_q)$.

3.2. *The queue-difference process.* Let

$$\Delta(t) \equiv q_2(t) - q_1(t), \quad t \ge 0.$$

As indicated in (3.1), at time T_1 we have $\Delta(T_1) = \kappa$. If $\dot{\Delta}(T_1) < 0$, then $\Delta(T_1+t) < \kappa$ for all t in some interval $(0, \epsilon]$ for $\epsilon > 0$. In that case, fluid from queue 2 stops flowing into pool 1. At some point $t_0 > T_1$ we may have that $-\Delta(t_0) = \kappa$, in which case sharing should begin with pool 2 helping queue 1, unless $z_{2,1}(t_0) > \tau$, which means that x will cross the sliding manifold $\mathbb{S}_{1,2}$ into $\mathbb{S}^+_{1,2}$. We now study the difference process over $[0, \Sigma_2)$. In terms of (3.6),

(3.11)

$$\Delta(t) = \Delta(0)e^{-\theta t} - \frac{1+\mu}{\theta}(1 - e^{-\theta t}) - \left(\frac{(1-\mu)(1 - z_{2,1}(0))}{1-\theta}\right)(e^{-\theta t} - e^{-t}) + \left(\frac{(1-\mu)\tau}{\mu - \theta}\right)(e^{-\theta t} - e^{-\mu t}).$$

LEMMA 3.1 (derivative of Δ over $\mathcal{I}_1$). *The function Δ in (3.11) has a negative derivative on $\mathcal{I}_1$ and is therefore strictly decreasing. In particular,*

$$\dot{\Delta}(t) = -\theta\Delta(t) + \Psi(t), \quad t \in \mathcal{I}_1, \tag{3.12}$$

where $\Delta(t) > 0$ and

$$\Psi(t) \equiv -(1+\mu) - (1-\mu)(1 - z_{2,1}(0))e^{-t} + (1-\mu)\tau e^{-\mu t} < 0, \quad t \in \mathcal{I}_1, \tag{3.13}$$

so that $\dot{\Delta}(t) < 0$ and $-\Psi_U \le \Psi(t) \le -\Psi_L$, where

$$0 < \Psi_L \equiv 2\mu - (1-\mu)(1-\tau) < 2 \equiv \Psi_U < \infty, \quad t \in \mathcal{I}_1. \tag{3.14}$$

PROOF. The expression for the derivative (prior to time T_1) follows immediately from (3.5). The function $\Psi(t)$ in (3.13) is strictly negative because $1+\mu > 1-\mu > (1-\mu)\tau e^{-\mu t}$ for all $t \geq 0$. □

We also have an explicit expression for the difference at time t in terms of its value at time 0. Specifically, (3.12) is a classic first-order ordinary differential equation, which is known to have the explicit solution

$$\Delta(t) = \Delta(0)e^{-\theta t} + e^{-\theta t}\int_0^t e^{\theta s}\Psi(s)\,ds, \quad t \in \mathcal{I}_1, \tag{3.15}$$

where $\Psi(s)$ is defined in *(3.13)* and is independent of $\Delta(0)$. Thus, $\Delta(t)$ is a strictly increasing function of the initial difference $\Delta(0) > 0$. In addition, $\Psi(s)$ and $\Delta(t)$ are increasing functions of $z_{2,1}(0)$ and τ. As a consequence, T_1 is strictly increasing function of $\Delta(0)$, $z_{2,1}(0)$ and τ. Moreover, for Ψ_L and Ψ_U in *(3.14)* and $t \in \mathcal{I}_1$,

$$\Delta(0)e^{-\theta t} - \Psi_U\left(\frac{1-e^{-\theta t}}{\theta}\right) \leq \Delta(t) \leq \Delta(0)e^{-\theta t} - \Psi_L\left(\frac{1-e^{-\theta t}}{\theta}\right). \tag{3.16}$$

From (3.8), we immediately obtain an expression for the derivative of the queue difference:

$$\dot{\Delta}(T_1+t) \;=\; -\theta\Delta(T_1+t) + Ae^{-\mu t}, \quad 0 \leq t \leq T_2, \tag{3.17}$$

where $\Delta(T_1) = \kappa$ and

$$A \equiv (1-\mu)(z_{1,2}(T_1) - z_{2,1}(T_1)) < 0. \tag{3.18}$$

Hence, $\dot{\Delta}(t) < 0$, so that $d_{2,1}(t) < 0$ ($q_2(t) < q_1(t) + \kappa$) for all $t \in \mathcal{I}_2$. Therefore, $\dot{\Delta}(t) < 0$ for all $t \in [0, \Sigma_2)$, so that Δ is strictly decreasing over that interval, implying that the time T_1 is well defined as the unique solution t to the equation $\Delta(t) = \kappa$.

Finally, (3.9) implies that the function $\Delta(t)$ can be expressed as

$$\Delta(T_1+t) \;=\; \kappa e^{-\theta t} + \Phi(t), \quad 0 \leq t \leq T_2, \tag{3.19}$$

where, for all $0 \leq t \leq T_2$,

$$\Phi(t) \equiv Ae^{-\theta t}\int_0^t e^{\theta s}e^{-\mu s}\,ds = A\left(\frac{e^{-\theta t} - e^{-\mu t}}{\mu - \theta}\right) < 0, \tag{3.20}$$

with $A < 0$ in *(3.18)*. In particular, $\Delta(T_1+t) < \kappa$ for all $t \in \mathcal{I}_2$, so that there is no active sharing in $\mathcal{I}_2$.

3.3. *Conditions for finiteness of the switching times.* From the definition of T_1 in (3.1) together with (3.16), we immediately get that $T_1 < \infty$. Given T_1, we can apply (3.7) to obtain an equation for T_2. If T_1 is sufficiently large so that $z_{2,1}(T_1) > \tau$, then

$$\begin{aligned} z_{2,1}(\Sigma_2) \equiv z_{2,1}(T_1 + T_2) &= z_{2,1}(T_1)e^{-\mu T_2} = [1 - (1 - z_{2,1}(0))e^{-T_1}]e^{-\mu T_2} \\ &= z_{1,2}(0) = \tau, \end{aligned}$$

where the last equality follows from the definition of T_2. As an immediate consequence of (3.1), we have explicit formulas for T_2:

$$T_2 = \frac{\log_e(z_{2,1}(T_1)/\tau)}{\mu} = \frac{\log_e([1 - (1 - z_{2,1}(0))e^{-T_1}]/\tau)}{\mu}. \tag{3.21}$$

It is easy to check whether $z_{2,1}(T_1) > \tau$ so that $T_2 > 0$; see (3.4) above. It suffices to have $e^{-T_1} < 1 - \tau$ or, equivalently, $T_1 > -\log_e(1 - \tau)$.

Combining (3.7) with (3.21) to obtain an expression for $z_{1,2}(\Sigma_2)$

$$z_{1,2}(\Sigma_2) = \tau e^{-\mu\Sigma_2}. \tag{3.22}$$

We can apply (3.9) to calculate $q_i(\Sigma_2)$ to verify that $q_i(\Sigma_2) > 0$ for $i = 1, 2$, ensuring that $\Sigma_q \geq \Sigma_2$. If $x(\Sigma_2)$ satisfies the conditions of $x(0)$ in Assumption 2 but with the labels of the processes reversed, then we can again apply (3.7) (with the labels reversed) to conclude that $T_3 < \infty$. If $T_3 > 0$, then T_4 satisfies a similar equation to (3.21), but with T_3 replacing T_1 and $z_{1,2}(T_3)$ replacing $z_{2,1}(T_1)$, provided that $z_{1,2}(T_3) > \tau$.

4. Qualitative analysis. Just as for the stochastic system, it is important to identify the possible equilibrium behavior of the fluid models, as well as its long-run behavior. We start with formally defining the relevant equilibria for our fluid model and then stating the main results regarding fluid model.

Recall that the state space of the fluid model is $\mathbb{S}$ in (2.7). For the general discussion regarding the long-run behavior of the system, we consider all the possible initial conditions, and therefore Assumption 2 is not enforced in this section. Specifically, any $\gamma \in \mathbb{S}$ is allowed to be an initial condition.

DEFINITION 4.1 (stationary point). *A point $x^* \in \mathbb{S}$ is stationary for (3.10) if $x(0) = x^*$ implies that $x(t) = x^*$ for all $t \geq 0$.*

DEFINITION 4.2 (periodic equilibrium). *A non-constant solution $u^* \equiv \{u^*(t) : t \geq 0\}$ to (3.10) is a periodic equilibrium, if there exists $T > 0$ such that $u^*(t + T) = u^*(t)$ for all $t \geq 0$. The smallest such T is called the period of u^*.*

Lyapunov stability of a stationary point. We will show that for any set of parameters, the fluid model in Definition 3.1 has a unique stationary point and that, in some cases, there also exists a unique periodic equilibrium. We will then study the stability properties of the fluid model. There are three types of stability notions corresponding to stationary points that are relevant for us.

For a stationary point x^*, let $\mathcal{S}_{x^*} \subseteq \mathbb{S}$ be the stability region of x^*, i.e., if $x(0) \in \mathcal{S}_{x^*}$, then $x(t) \to x^*$ as $t \to \infty$. Note that, by the definition of x^*, $\mathcal{S}_{x^*}$ is not empty because it contains x^*.

DEFINITION 4.3 (Lyapunov stability). *A stationary point x^* is said to be*

- ***unstable**, if $\mathcal{S}_{x^*} = \{x^*\}$;*
- ***asymptotically stable**, if $\mathcal{S}_{x^*}$ contains an open neighborhood of x^*;*
- ***globally asymptotically stable**, if $\mathcal{S}_{x^*} = \mathbb{S}$.*

We note that for our system with the state space $\mathbb{S}$ in (2.7), subsets of $\mathbb{S} \subsetneq \mathbb{R}_6$ are considered open in the relative topology induced on $\mathbb{S}$ by the topology of $\mathbb{R}_6$. In particular, open subsets can contain points on the boundary of $\mathbb{S}$ in $\mathbb{R}_6$.

Stability of a periodic equilibrium. When a periodic equilibrium u^* exists, it is possible for the fluid model to oscillate indefinitely, at least when the initial condition is taken to be on the periodic equilibrium trajectory. However, we would like to know if the periodic equilibrium is also asymptotically stable in some sense, namely, if there exists a set $\mathcal{S}_{u^*} \subseteq \mathbb{S}$ such that, if $x(0) \in \mathcal{S}_{u^*}$, then $x(t)$ converges to the periodic equilibrium. We note that convergence to periodic equilibrium cannot hold in the Lyapunov sense, as in Definition 4.3, because there would typically be a time shift between the converging solution and the periodic-equilibrium solution. We therefore say that a solution x converges to a periodic equilibrium u^* if its image "spirals" toward the image of u^* as time increases. (By spiraling we mean that the image of x keeps moving in the direction of u^* and gets closer to it as time increases; see Lemma D.4 in the appendix.)

Consider a switching dynamical system $\dot{x} = f_\sigma(x)$ (not necessarily (3.10)). The standard way of proving that a periodic equilibrium u^* (assuming one exists) with period T is stable, is to consider the intersection point $\tilde{u}$ of u^* with a switching surface $\mathcal{M}$, and show that any trajectory x that is initialized on $\mathcal{M}$ sufficiently close to $\tilde{u}$, will reach $\mathcal{M}$ again after a time that is approximately equal to the period T of u^*. If, in addition, the intersections

of x with $\mathcal{M}$ converge to $\tilde{u}$, then u^* is asymptotically stable; see, e.g., page 121 in [40].

To rigorously define the above asymptotic stability notion, and show that it indeed implies the "spiraling motion" of solutions that are initialized sufficiently close to a periodic equilibrium, we first make a simple observation: When there are $N > 1$ switching surfaces $\mathcal{M}_i$, $1 < i \leq N$, that are intersected by a *stable* periodic equilibrium u^*, the intersections of x with $\mathcal{M}_i$, as well as the values of x at those intersection points, will converge to the intersection points of u^* with $\mathcal{M}_i$ and the values of u^* at these epochs, respectively, for each $i \leq N$. Since this is the case for our system, we define asymptotic stability in term of all four switching surfaces and the corresponding switching times. To avoid introducing more notation, the definition is given for our system directly.

Let $\mathcal{P}_{u^*}$ denote the image of a periodic equilibrium u^* having period T;

$$\mathcal{P}_{u^*} \equiv \{\gamma \in \mathbb{S} : \gamma = u^*(t),\ 0 \leq t < T\}.$$

Since $u^*(0) = u^*(T)$, the set $\mathcal{P}_{u^*}$ is an *invariant* set, namely, if $y_0 \in \mathcal{P}_{u^*}$ and y is the unique solution to $\dot{y} = f_\sigma(y)$ in (3.10) with initial condition $y(0) = y_0$, then $y(t) \in \mathcal{P}_{u^*}$ for all $t > 0$.

Let x be a solution to (3.10) with $x(0) \notin \mathcal{P}_{u^*}$ and $\Sigma_q = \infty$ (so that x oscillates indefinitely; we will show in Theorem 5.5 below that such solutions exist). Note that if x is an oscillating solution to (3.10), then there exists a $t_1 \geq 0$ such that $x(t_1)$ satisfies the conditions in Assumption 2. Due to the time-homogeneity of x we can restart the ODE at the first time $t_1 \geq 0$ for which $x(t_1)$ satisfies Assumption 2 by taking $x(0) = x(t_1)$. Then the solution $\{x(t) : -t_1 \leq t < \infty\}$ satisfies Assumption 2 at time 0.

For T_i and Σ_i in (3.1) and (3.2), let $T_i^{(k)}$ and $\Sigma_i^{(k)}$ be the value of holding time T_i and switching time Σ_i, respectively, in the k^{th} cycle of x, where

$$\Sigma_0^{(1)} \equiv t_1 \text{ (so that } x(\Sigma_0^{(1)}) \equiv x(0) \text{ by definition) and } \Sigma_0^{(k+1)} \equiv \Sigma_4^{(k)},\ k \geq 1.$$

Let T_j^* denote holding time j, $1 \leq j \leq 4$, and $\Sigma_i^{*(k)}$ denote switching time i, $0 \leq i \leq 4$, in the k^{th} cycle of a periodic equilibrium u^*, with $\Sigma_0^{*(0)} \equiv 0$ and $\Sigma_0^{*(k+1)} \equiv \Sigma_4^{*(k)}$, $k \geq 1$. Similarly, for an oscillating solution x, let $T_j^{(k)}$, denote holding time j, $1 \leq j \leq 4$, and $\Sigma_i^{(k)}$ denote switching time i, $0 \leq i \leq 4$, in the k^{th} cycle of x, $k \geq 1$, where $\Sigma_0^{(0)} \equiv 0$ and $\Sigma_0^{(k+1)} \equiv \Sigma_4^{(k)}$, $k \geq 1$.

DEFINITION 4.4 (asymptotically stable periodic equilibrium). *A periodic equilibrium u^* having period T is said to be asymptotically stable if there*

exists an open subset $\mathcal{S}_{u^*}$ *of* $\mathbb{S}$ *which contains* $\mathcal{P}_{u^*}$ *such that, if* $x(0) \in \mathcal{S}_{u^*}$, *then for* $1 \le i \le 4$ *and any* $t > 0$,

$$\lim_{k\to\infty} T_i^{(k)} = T_i^* \quad \textit{and} \quad \lim_{k\to\infty} \sup_{0\le s\le t} \|x(\Sigma_0^{(k)} + s) - u^*(\Sigma_0^{*(k)} + s)\| = 0. \tag{4.1}$$

5. Asymptotic behavior of the fluid model. In this section we establish results about the asymptotic behavior of the switching fluid model in (3.10). We show that there always is the underloaded stationary point equilibrium, to which the fluid model converges if it does not oscillate indefinitely. We show that there exists an overloaded periodic equilibrium if it oscillates indefinitely, and provide sufficient conditions for endless oscillations. For the discussion of equilibria, we no longer assume initial conditions in Assumption 2; we allow arbitrary initial conditions in the state space $\mathbb{S}$. We also consider the system after time Σ_q in (3.3). The proofs of all the results in this section are relegated to the Appendix.

5.1. *Existence and asymptotic stability of a unique stationary point.* If there is no sharing actively taking place on an interval $[0, T]$, then the stochastic system decomposes into two independent $M/M/n+M$ (Erlang-A) queuing systems. Let $Y_i^n(t) := Q_i^n(t) + Z_{i,i}^n(t)$ denote the total number of customers in each of these systems and $\bar{Y}_i^n := Y_i^n/n$, $i = 1, 2$. Then the fluid model for $\bar{Y}^n$ in the symmetric case we consider is the solution of the ODE

$$\dot{y}_i = \lambda - \mu(1 \wedge y_i) - \theta(y_i - 1)^+, \quad i = 1, 2,$$

where $a^+ \equiv \max\{a, 0\}$. In this case we have the following elementary, but important, result.

THEOREM 5.1. *If* $q_i(0) \le \kappa$, *then no sharing will ever begin in the fluid model and* $x(t) \to x_0^*$ *as* $t \to \infty$, *where*

$$x_0^* \equiv (q_1^*, q_2^*, z_{1,1}^*, z_{1,2}^*, z_{2,1}^*, z_{2,2}^*) = (0, 0, \lambda, 0, 0, \lambda). \tag{5.1}$$

Hence, x_0^* *is an asymptotically stable stationary point.*

REMARK 5.1. Having x_0^* in (5.1) be an asymptotically stable stationary point depends critically on the assumption that $\kappa > 0$. If, instead, $\kappa = 0$, then it is possible for x_0^* to be an unstable stationary point, so that x oscillates indefinitely for any initial condition $x(0) \ne x_0^*$. Instability of x_0^* has important consequences for the stochastic system X^n, since stochastic fluctuations may trigger undesirable sharing even if the system is initialized at the neighborhood of x_0^*. Therefore, stochastic fluctuations can quickly

lead to *fluid-scaled fluctuations*, namely, to an oscillatory behavior. See the simulations in §7.4 below. The moral is that there is a need to ensure that the activation thresholds in the (finite) stochastic system are large enough to be considered positive in fluid scale. The size of the stochastic fluctuations of critically-loaded pools with no sharing can be estimated from the established heavy-traffic limit approximations for the Erlang-A model in [13].

Ideally, x_0^* in (5.1) would be a globally asymptotically stable stationary point for the fluid model, since the system is underloaded ($\lambda < 1$) and we want no sharing to take place, and indeed that will be the case with appropriate controls. However, here we are interested in fluid models with poorly chosen controls. Then solutions to (3.10) need not converge to x_0^*, so that $\mathcal{S}_{x_0^*}^c \neq \phi$, where, for a set A, A^c denotes the complement of A and ϕ denotes the empty set.

Let $\mathbb{S}^* := \{\gamma^* \in \mathbb{S} : \gamma^* \text{ is a stationary point}\}$.

THEOREM 5.2. *$\mathbb{S}^* = \{x_0^*\}$ for x_0^* in (5.1); i.e., x_0^* is the unique stationary point of the switching fluid model.*

Due to Theorem 5.2, we can refer to x_0^* in (5.1) as the *stationary point with no sharing*, or simply as the stationary point.

5.2. *Only two possibilities.* We now show that there are only two possibilities for the asymptotic behavior. Let $\mathcal{O} \subset \mathbb{S}$ be the set of points such that, if $x(0) \in \mathcal{O}$, then the solution x to (3.10) switches infinitely often as $t \to \infty$, i.e., it oscillates indefinitely.

THEOREM 5.3. *$\mathcal{O}^c = \mathcal{S}_{x_0^*}$ for x_0^* in (5.1); i.e., if $x(0) \in \mathcal{O}^c$, then $x(t) \to x_0^*$ as $t \to \infty$.*

5.3. *Existence of a periodic equilibrium.* Theorem 5.3 shows that a solution x to (3.10) either converges to x_0^* or oscillates indefinitely. We now consider what happens if the solution oscillates indefinitely.

THEOREM 5.4. *If $\mathcal{O} \neq \phi$, then there exists a periodic equilibrium $u^* \equiv \{u^*(t) : t \geq 0\}$ to (3.10). In particular, if $\mathcal{O} \neq \phi$, then there exists a initial state vector $x(0)$ satisfying Assumption 2 such that $x(0) \in \mathcal{O}$ and, for that $x(0)$, $\Sigma_q > \Sigma_2$ and $(q_1(\Sigma_4), q_2(\Sigma_4), z_{2,1}(\Sigma_4)) = (q_1(0), q_2(0), z_{2,1}(0))$, which implies that $T_3 = T_1$, $T_4 = T_2$, so that $\Sigma_4 = 2\Sigma_2$,*

(5.2)
$$\begin{aligned}(q_1(\Sigma_4), q_2(\Sigma_4), z_{2,1}(\Sigma_4)) &= (q_1(2(T_1+T_2)), q_2(2(T_1+T_2)), z_{2,1}(2(T_1+T_2)))\\ &= (q_2(T_1+T_2), q_1(T_1+T_2), z_{1,2}(T_1+T_2))\\ &= (q_1(0), q_2(0), z_{2,1}(0)).\end{aligned}$$

It is important that the condition in Theorem 5.4 can be satisfied, as the following theorem shows.

THEOREM 5.5. *There exist parameter values for* (2.4) *and initial conditions satisfying Assumption 2 for which* $\mathcal{O} \neq \phi$.

5.4. *Conjectured bi-stability.* Recall that $\mathcal{S}_{x_0^*}$ is the stability set of x_0^* in Definition 4.3 and $\mathcal{S}_{u^*}$ denotes the stability set of the periodic equilibrium u^*, when it exists, in Definition 4.4. By Theorem 5.3, $\mathcal{S}_{x_0^*} = \mathcal{O}^c$ (the complement of $\mathcal{O}$), so that any fluid solution that *does not* oscillate indefinitely must converge to x_0^*, and it clearly holds that $\mathcal{S}_{u^*} \subseteq \mathcal{O}$. We conjecture that $\mathcal{S}_{u^*} \supseteq \mathcal{O}$ as well, so that $\mathcal{S}_{u^*} = \mathcal{O}$. Formally,

CONJECTURE 5.1. *If* $x(0) \in \mathcal{O}$, *then there exists a unique periodic equilibrium* u^* *and* x *converges to* u^* *as in* (4.1). *Therefore,* $\mathcal{S}_{x_0^*} \cup \mathcal{S}_{u^*} = \mathbb{S}$, *namely the fluid model is bi-stable with all fluid trajectories converging to one of the two equilibria as* $t \to \infty$.

Extensive numerical trials, some of which are presented in §7 below, indicate that Conjecture 5.1 holds. Moreover, we next derive an approximating system to (3.10) which is shown to be bi-stable.

6. Approximating dynamical system. Since we were unable to fully characterize the asymptotic behavior of our initial fluid model, we now develop an approximating fluid model that can be analyzed more easily; i.e., for which we can establish bistability and calculate the two equilibria. The approximating system is easier to analyze because it is essentially a one-dimensional system at the switching times. However, there are discontinuities at some of the switching times, so the approximating fluid model is a dynamical system with jumps (alternatively, it can be represented as a hybrid system with jumps); see [38] and [42]. The latter reference provides a general framework for defining and analyzing solutions for dynamical systems with jumps (see §1.5 of [42]), but the relative simplicity of our approximation obviates the need for a general theory. Numerical examples confirm that the approximating system serves as a useful approximation for the original fluid model, allowing us to rapidly compute a periodic equilibrium.

The approximation is obtained in five steps: First, we approximate the solution x to (3.10) by a solution x^a to

$$\dot{x}^a = f_\sigma(x^a, \theta^a, \tau^a), \tag{6.1}$$

for a given initial condition $x^a(0)$, where we supplement the argument x^a of f_σ in (3.10) by the abandonment rate θ^a and the control parameter τ^a of the approximating system. Second, we assume that there is no abandonment, i.e., we let $\theta^a = 0$. Third, approximate τ by 0 on the first and third subintervals, i.e.,

$$\tau^a \equiv \begin{cases} 0 & \text{for } 0 \le t < \Sigma_1^a \quad \text{and} \quad \Sigma_2^a \le t < \Sigma_3^a \\ \tau & \text{for } \Sigma_1^a \le t < \Sigma_2^a \quad \text{and} \quad \Sigma_3^a \le t < \Sigma_4^a, \end{cases} \tag{6.2}$$

where the switching times Σ_i^a are defined analogously to (3.2), and are formally defined in (6.5) below. Fourth, we let the initial condition for the approximating system be defined by

$$x^a(0) = \lim_{\tau \to 0} x(0), \quad \text{so that } z_{1,2}^a(0) = z_{2,1}^a(0) = 0, \tag{6.3}$$

where $x(0)$ is the initial condition in Assumption 2. Fifth, and finally, we primarily focus on the three-dimensional function $x_3^a \equiv (\Delta^a, z_{1,2}^a, z_{2,1}^a)$ that approximates the three-dimensional function $x_3 \equiv (\Delta, z_{1,2}, z_{2,1})$ obtained from (3.10), ignoring the queue lengths. We will be assuming that the queue lengths remain positive, which can be checked at the end. In general, our analysis is valid until a queue length becomes 0. First, we focus on the difference function because it is possible to do so and still have a bonafide dynamical system, which is easier to analyze. Second, we are motivated to ignore the queue lengths because we have less control over them without abandonment; e.g., they can easily explode (diverge to infinity). However, we will also state results for the full six-dimensional approximation x^a.

Since the approximating queue lengths q_1^a and q_2^a can obtain any nonnegative value, the full state space $\mathbb{S} \equiv [0, \lambda/\theta]^2 \times [0,1]^4$ of the solutions to (3.10) is replaced with $\mathbb{S}^a \equiv [0,\infty)^2 \times [0,1]^4$. Indeed $\mathbb{S}^a$ is obtained from $\mathbb{S}$ directly because $\lambda/\theta \to \infty$ as $\theta \to 0$. The state space of x_3^a is a-priori $[0,\infty) \times [0,1]^4$, but we will show below that Δ is bounded from above.

Paralleling (3.1), the switching and holding times, and the intervals between switching times, are defined via

$$\begin{aligned} T_1^a &\equiv \inf\{t \ge 0 : q_2^a(t) - q_1^a(t) \le \kappa\} \\ T_2^a &\equiv \inf\{t \ge 0 : z_{2,1}^a(\Sigma_1^a + t) \le \tau\}, \\ T_3^a &\equiv \inf\{t \ge 0 : q_1^a(\Sigma_2^a + t) - q_2^a(\Sigma_2^a + t) \le \kappa\} \\ T_4^a &\equiv \inf\{t \ge 0 : z_{1,2}^a(\Sigma_3^a + t) \le \tau\}, \end{aligned} \tag{6.4}$$

where, with $T_0^a \equiv \Sigma_0^a \equiv 0$,

$$\Sigma_k^a \equiv \sum_{i=0}^{k} T_i^a \quad \text{and} \quad \mathcal{I}_i^a \equiv [\Sigma_{i-1}^a, \Sigma_i^a), \quad k = 1, 2, 3, 4. \tag{6.5}$$

Paralleling (3.3), we let

$$\Sigma_q^a \equiv \inf\{t > 0 : q_1^a(t) \wedge q_2^a(t) = 0\}. \tag{6.6}$$

Our analysis will be valid for the full six-dimensional system on the interval $[0, \Sigma_q^a]$, but we will not examine Σ_q^a until the end. In particular, we will show that the system quickly converges to the (unique) periodic equilibrium, when it exists, for any initial condition that is associated with an oscillating solution. We can therefore initialize the queues (which are unbounded) at large values so that there is no time for them to reach 0 by the time convergence to the periodic equilibrium is observed.

In examples we see that the approximating system approximates our original system very well when the parameters θ and τ are suitably small. For this approximating system, we establish the following two results (Theorems 6.1 and 6.2). Let $\Sigma_4^{a,(k)}$ and $\Delta^{a,(k)}$ be the values of the k^{th} iteration, where we apply the approximation above in the k^{th} subinterval after making $\Sigma_4^{a,(k-1)}$ equal to time 0.

The first main result regarding the approximating system, Theorem 6.1 below, considers the case in which the approximating system converges to its unique fixed point.

THEOREM 6.1. *Consider the approximating system defined in* (6.1)–(6.6).

(a) The unique stationary point x_0^* in (5.1) *for the fluid model in* §3 *is also the unique stationary point in $\mathbb{R}^6$ for the approximating system.*

(b) If $\Delta^a(0) \le \kappa$ or if $\Delta^{a,(k)}(0) \le \kappa$ for some $k \ge 1$, then $x^a(t) \to x_0^*$ in $\mathbb{R}^6$ for x_0^* in (5.1).

(c) Whenever $x^a(t) \to x_0^*$ in $\mathbb{R}^6$ for x_0^* in (5.1), *$x_3^a(t) = (0, 0, 0)$ for all sufficiently large t, namely, convergence to x_0^* occurs in finite time.*

The next theorem considers the case in which the approximating system possesses a periodic equilibrium, in addition to its unique stationary point x_0^*.

THEOREM 6.2. *Consider the approximating system defined in* (6.1)–(6.6).

(a) *If* $\Delta^{a,(k)}(0) > \kappa$ *for all* k, *then* $\Delta^{a,(k)}(0) \to \Delta^{a,(\infty)}(0) \in [\kappa + \epsilon^a_\kappa, (1-\mu)(1-\tau)/\mu]$ *as* $k \to \infty$, *where* $\epsilon^a_\kappa \equiv -\log(1-\tau) > 0$.

(b) *If the condition in part (a) holds, and if* $\Sigma^a_q = \infty$, *then (i) there exists a unique periodic equilibrium* u^{a*}_3 *to the three-dimensional approximating system and (ii) the approximating system is bistable: There are initial conditions for which* $x^a(t) \to x^*_0$ *in* $\mathbb{R}^6$ *for* x^*_0 *in* (5.1) *(which may include having* $\Sigma^a_q < \infty$*); there are other initial conditions for which* $\Sigma^a_q = \infty$ *and* $x^a(t)$ *fails to converge in* $\mathbb{R}^6$ *in the usual sense of pointwise convergence, but* $x^a_3(t) \to u^{a*}_3$ *in* $\mathbb{R}^3$ *in the sense of Definition 4.4; and there are no other possibilities.*

(c) *For any given pair of control parameters* (κ, τ), *there exists* $\mu^* \equiv \mu^*(\kappa, \tau)$ *such that, for any service rate* $\mu \in (0, \mu^*)$, *the condition in part (a) holds with* $\Delta^{a,(\infty)}(0) > \kappa$, *so that the conclusions of part (b) hold, provided that* $\Sigma^a_q = \infty$.

In particular, by Theorem 6.2 (b), when a periodic equilibrium exists to the approximating system, then the system is bi-stable; each solution must converge to one of the two equilibria x^*_0 or u^{a*}_3. Otherwise, all solutions converge to x^*_0 (which is therefore a globally asymptotically stable stationary point in this case).

The condition $\Sigma^a_q = \infty$ in part (b) of Theorem 6.2 is easy to check directly by solving the simple equations for the full six-dimensional equation (6.1). However, in Appendix F we show that, whether or not this condition holds can be determined a posteriori by a simple calculation that depends only on the periodic equilibrium, and does not depend on the transient behavior of the fluid model.

In §6.1 and §6.2 we derive the solution to the approximating system over the first and second intervals, $[0, \Sigma^a_1)$ and $[\Sigma^a_1, \Sigma^a_2)$, respectively. In §6.3 we construct the solution after Σ^a_2. In §6.4 we consider a simple heuristic to provide an approximate explicit formula for the switching time T^a_1 to facilitate computations.

All the results in this section are proved in §D in the appendix. Furthermore, in §F we show how to apply the explicit formula in §6.4 to determine if there will be congestion collapse. We establish conditions for a stronger geometric rate of convergence and exponential stability in §H.

6.1. *The approximation over the first interval* $\mathcal{I}^a_1 = [0, \Sigma^a_1)$. The ODE's for x^a over $[0, \Sigma^a_1)$ are just as in (3.4)–(3.5), but with $\theta = \tau = 0$. Just as in §3.1.1, q^a_1 is increasing while $q^a_2 \geq q^a_1 + \kappa$, so $\Sigma^a_q > \Sigma^a_1$.

It follows from (3.4) that, for $x^a(0)$ in (6.3) and $0 \le t < \Sigma_1^a$,

$$z_{1,2}(t) = 0 \text{ and } z_{2,1}(t) = 1 - e^{-t}, \text{ so that } z_{1,1}(t) = e^{-t} \text{ and } z_{2,2}(t) = 1. \tag{6.7}$$

The value of T_1^a is determined by the process $\Delta^a \equiv q_2^a - q_1^a$, approximating the corresponding difference process Δ. Taking $\theta = \tau = 0$ and $z_{2,1}(0) = 0$ in (3.12)–(3.13), we have, for $0 \le t < \Sigma_1^a$, $\dot{\Delta}^a(t) = -(1+\mu) - (1-\mu)e^{-t}$, so that

$$\Delta^a(t) = \Delta^a(0) - (1+\mu)t + (1-\mu)(1-e^{-t}). \tag{6.8}$$

Since $\Delta^a(T_1^a) = \kappa$ by definition, it follows that

$$T_1^a = \frac{\Delta^a(0) - 1 + \mu - \kappa}{1+\mu} + \frac{1-\mu}{1+\mu} e^{-T_1^a}. \tag{6.9}$$

LEMMA 6.1. *For any fixed $\Delta^a(0) > \kappa$ there exists a unique $T_1^a > 0$ satisfying (6.9). Furthermore, T_1^a is strictly increasing in $\Delta^a(0)$.*

It follows from (6.7) that for $\Sigma_1^a \equiv T_1^a$,

$$x_3^a(\Sigma_1) = (\kappa, 0, 1 - e^{-T_1^a}), \tag{6.10}$$

which is well-defined by Lemma 6.1.

6.2. *The approximation over the second interval $\mathcal{I}_2^a = [\Sigma_1^a, \Sigma_2^a)$.* The equations for the service process over $[\Sigma_1^a, \Sigma_2^a)$ are obtained from (3.7), but with T_1^a replacing T_1 and $z_{i,j}^a(T_1^a)$ replacing $z_{i,j}(T_1)$, $i, j = 1, 2$. As in §3.1.2, it is possible to have $\Sigma_1^a < \Sigma_q^a \le \Sigma_2^a$, but we do not check that now.

Since the process $z_{1,2}$ in (3.7) keeps decreasing and $z_{1,2}^a(T_1^a) = 0$, it follows from (6.10) and (3.7) that

$$z_{1,2}^a(T_1^a + t) = 0 \quad \text{and} \quad z_{2,1}^a(T_1^a + t) = (1 - e^{-T_1^a})e^{-\mu t}, \quad 0 \le t < T_2^a. \tag{6.11}$$

Taking $\theta \downarrow 0$ and inserting the values of $z_{1,2}^a(T_1^a)$ and $z_{2,1}^a(T_1^a)$ from (6.10) in (3.17), we see that

$$\dot{\Delta}^a(\Sigma_1^a + t) = -z_{2,1}^a(T_1^a)(1-\mu)e^{-\mu t} = -(1 - e^{-T_1^a})(1-\mu)e^{-\mu t}, \tag{6.12}$$

for $0 \le t < T_2^a$, where $\Delta^a(\Sigma_1^a) = \kappa$.

By (6.4), T_2^a is the first time after Σ_1^a that $z_{2,1}^a$ hits τ, so that, paralleling (3.21),

$$T_2^a = \frac{\log(z_{2,1}^a(T_1^a)/\tau)}{\mu} = \frac{\log((1 - e^{-T_1^a})/\tau)}{\mu}. \tag{6.13}$$

Clearly, if $\tau \downarrow 0$ then $T_2^a \to \infty$, which is why we cannot replace τ with 0 over the second interval $[\Sigma_1^a, \Sigma_2^a)$.

Inserting the value of T_2^a into the solution to (6.12) we obtain

$$\Delta^a(\Sigma_2^a-) = \kappa - \frac{z_{2,1}(T_1^a)(1-\mu)}{\mu}\left(1 - \frac{\tau}{z_{2,1}^a(T_1^a)}\right) = \kappa - \frac{1-\mu}{\mu}(1 - e^{-T_1^a} - \tau),$$

where $y(t-) \equiv \lim_{s\uparrow t} y(s)$ denotes the left limit at time t of a function y. Hence,

$$x_3^a(\Sigma_2^a-) = \left(\kappa - \frac{1-\mu}{\mu}(1 - e^{-T_1^a} - \tau), 0, \tau\right). \tag{6.14}$$

6.3. *Continuing beyond* Σ_2^a. As before, we can use the symmetry of x_3^a and take $x_3^a(\Sigma_2^a)$ to be the "initial condition" by reversing the labels. This means that, as in (6.3), we take $\tau \downarrow 0$ in $x_3^a(\Sigma_2^a)$. It follows immediately from (6.14) that $\lim_{\tau\downarrow 0} x_3^a(\Sigma_2^a) \neq x_3^a(\Sigma_2^a-)$. Hence, the approximation x_3^a, and therefore x^a, has a jump at time Σ_2^a, since the values of $\Delta^a(\Sigma_2^a-)$ and $z_{2,1}(\Sigma_2^a-)$ both depend on τ. However, we can easily avoid having jumps in the process Δ^a, which we want to avoid because it causes ambiguities about the behavior of the queues at the jump times. To that end, we simply define

$$\Delta^a(\Sigma_2^a) \equiv \Delta^a(\Sigma_2^a-) = \kappa - \frac{1-\mu}{\mu}(1 - e^{-T_1^a} - \tau) \quad \text{and}$$
$$z_{2,1}(\Sigma_2^a) = \lim_{\tau\downarrow 0} z_{2,1}(\Sigma_2^a) = 0,$$

so that we have

$$x_3^a(\Sigma_2^a) = \left(\kappa - \frac{1-\mu}{\mu}(1 - e^{-T_1^a} - \tau), 0, 0\right). \tag{6.15}$$

As a consequence, only $z_{2,1}$ jumps at the second switching time Σ_2^a. That discontinuity makes our fluid model a switching dynamical system with jumps, as mentioned at the beginning of the section.

If $\Delta^a(\Sigma_2^a) > \kappa$, then $T_3^a > 0$, and paralleling (6.9) and Lemma 6.1, T_3^a is the unique strictly positive solution to

$$T_3^a = \frac{\Delta^a(\Sigma_2^a) - 1 + \mu - \kappa}{1+\mu} + \frac{1-\mu}{1+\mu}e^{-T_3^a}.$$

Furthermore, paralleling (6.13), $T_4^a = \frac{1}{\mu}\log((1 - e^{-T_3^a})/\tau)$, so that

$$\Delta^a(\Sigma_4^a-) = \frac{1-\mu}{\mu}(1-e^{-T_3^a} - \tau) - \kappa, \quad z_{1,2}^a(\Sigma_4^a-) = \tau \quad \text{and} \quad z_{2,1}^a(\Sigma_4^a-) = 0.$$

If $\Delta^a(\Sigma_4^a-) > \kappa$ we define

$$\Delta^a(\Sigma_4^a) \equiv \Delta^a(\Sigma_4^a-) \quad \text{and} \quad z_{1,2}^a(\Sigma_4^a) = \lim_{\tau \downarrow 0} z_{1,2}^a(\Sigma_4^a-) = 0$$

and start over.

The preceding shows that, just as for the original system, we can exploit the symmetry of the model and consider only the half cycle $[0, \Sigma_2^a)$. In particular, for a given initial condition $\Delta^a(0)$ we solve up to time Σ_2^a and take

$$-x_3^a(\Sigma_2^a) = \left(\frac{1-\mu}{\mu}(1 - e^{-T_1^a} - \tau) - \kappa, 0, 0\right) \tag{6.16}$$

to be a new initial condition to solve beyond time Σ_2^a. It immediately follows that

LEMMA 6.2. *Δ^a is bounded over $[0, \Sigma_q^a)$. In particular, if $\Sigma_4^a < \Sigma_q^a$, then $\Delta^a(\Sigma_4^a) < \Delta_{bd}^a \equiv \frac{1-\mu}{\mu}(1-\tau)$.*

It is significant that at the switching times, x_3^a depends only on the known control parameters (κ, τ) and the one unknown T_1^a. Therefore, *the approximating system is reduced to an essentially one-dimensional system at the switching times.*

The approximating three-dimensional system. From the above, $x_3^a = (\Delta^a, z_{1,2}, z_{2,1})$ is the unique solution over $[0, \Sigma_q^a)$, for Σ_q^a in (6.6), to

$$\dot{x}_3^a = f^3_{\sigma(x_3^a)}(x_3^a, \theta, \tau^a) = f^3_{\sigma(x_3^a)}(x_3^a, 0, \tau^a), \quad \sigma(x_3^a) = 1, 2, 3, 4, \tag{6.17}$$

with initial condition (6.3) and τ^a in (6.2), where f_1^3 is defined in (3.4) and (6.8), f_2^3 is defined in (3.7) and (6.12), f_3^3 satisfies the equations of f_1^3, but with the labels reversed, and f_4^3 satisfies the equations of f_2^3, with the labels of the processes reversed.

6.4. *A simple heuristic approximation for computation.* The approximating system we have developed in this section has been useful to estalbish the strong theoretical results in Theoreem 6.1, which supports what we see for the original system in numerical examples. However, it is still not easy to compute the periodic equilibrium of the approximating system. We must either numerically solve the ODE's or numerically solve for T_1^a in (6.9) in order to evaluate the values of x^a at the switching times. Hence, in the present section we develop a simple heuristic approximation for T_1^a in (6.9).

In particular, our approximation is obtained by simply omitting the second exponential term on the right in (6.9), so that

$$T_1^a \approx \frac{\Delta - 1 + \mu - \kappa}{1 + \mu}. \tag{6.18}$$

Approximation (6.18) can be justified by observing that equation (6.9) can be expressed abstractly as $T_1^a = A + Be^{-T_1^a}$ for $A > 0$ and $0 < B < 1$. Since $T_1^a > A$ and $T_1^a - A < Be - A$, $T_1^a \approx A$ whenever B is suitably small or A is suitably large. In particular, the error is asymptotically negligible as A increases. We remark that approximation (6.18) also coincides with $-\log(\xi)\ \xi \equiv \xi(\Delta)$ in (D.7), which can provide another way to derive the approximation. We can combine (6.10) and (6.18) ot obtain an associated approximation for $z_{2,1}^a(T_1^a)$.

With this heuristic approximation for $z_{2,1}^a(T_1^a)$, we have by (6.13) that

$$T_2^a \approx \frac{\log\left((1-\xi)/\tau\right)}{\mu}, \tag{6.19}$$

so that (6.14) and (6.15) are respectively approximated by

$$\begin{aligned} x^a(\Sigma_2^a-) &\approx \left(\kappa - \frac{1-\mu}{\mu}(1-\xi-\tau), 0, \tau\right) \\ \text{and} \quad x^a(\Sigma_2^a) &\approx \left(-\kappa + \frac{1-\mu}{\mu}(1-\xi-\tau), 0, 0\right), \end{aligned} \tag{6.20}$$

and $x^a(\Sigma_2^a)$ serves as the initial condition for the following cycle.

We can use this heuristic approximation to approximate the values of the fluid model at the switching times, using an iterative algorithm, which is described in §D.2.2. Furthermore, in §F we explain how this heuristic can be employed to evaluate whether a periodic equilibrium exists. A numerical example comparing the solution to the approximating system to the original fluid solution is presented in §7.1 below.

7. Numerical examples. In this section we report the results of numerical experiments based on numerical algorithms (numerical solutions of the dynamical systems) and simulations. Throughout this section we consider symmetric systems with parameters as in (2.4). In all our examples, $\lambda = 0.98$, $\tau = 0.01$ and $\kappa = 0.1$, but we vary the parameters θ and μ. The initial condition in the numerical examples is taken in accordance with Assumption 2.

We emphasize at the outset that μ in our numerical examples is taken to be extremely small. (We also consider systems with no abandonment, or

with very small abandonment rate, but this is prevalent in modeling.) However, as our simulation experiments below demonstrate, the oscillating fluid models for systems with extreme parameters suggest possible bad oscillatory dynamics in systems with more realistic parameters. In these more realistic setting the behavior cannot be predicted analytically, since the stochastic system is too complicated. Moreover, oscillations may even be overlooked in practice, because sufficient abandonment keep the queues relatively small, so that congestion collapse may fail to be noticed. Thus, we obtain important practical insights by rigorously studying extreme cases.

The rest of this section is organized as follows. In §7.1 we consider a system with no abandonment ($\theta = 0$) and compare the results to the heuristic approximating model in §6.4. We consider a similar system in §7.2 but increase μ to show that x_0^* is globally asymptotically stable, thus showing the dependence on μ of the long-run behavior of the fluid model, as was established in §6. We add abandonment in §7.3 in comparison to the system in §7.1 to numerically support the reasoning for the development of the approximating system in §6. Finally, in §7.4 we present simulations of stochastic systems for which the fluid limit has no oscillatory solutions, and show that stochasticity may lead to substantial oscillations.

7.1. *A system with no abandonment.* We start with a system that has no abandonment, i.e., $\theta = 0$. The other parameters are $\lambda = 0.98$, $\tau = 0.01$, $\kappa = 0.1$ and $\mu = 0.1$. The initial condition is $q_1(0) = 1$ and $q_2(0) = 1.2$, so that $d_{2,1}(0) = 0.2$ and $\Delta(0) = 0.1$. We further take $z_{1,2}(0) = \tau$ and $z_{2,1}(0) = \tau/2 = 0.005$.

The time-dependent behavior of Δ is shown in Figure 4, whereas Figure 5 plots the image of $(z_{2,1}, \Delta)$ (with time suppressed). As can be easily seen from Figure 4, there are ten full cycles plotted in this example. However, there are four loops visible in Figure 5, with each loop being a full cycle, where a full cycle begins at a time t_0 when $z_{1,2}(t_0)$ hits τ from above, such that Assumption 2 is satisfied at that hitting time. In this example, the two variables $(\Delta, z_{2,1})$ spiral outward to the periodic equilibrium, namely, the first cycle is the inner (smallest) loop, the second cycle is the second smallest loop, etc. The fact that only four cycles are clearly visible in Figure 5 suggests that convergence to the periodic equilibrium is extremely fast in terms of the number of periods. The fast convergence is also visible by in Figure 4 itself. See §H for theoretical support.

Of course, the stability of $(\Delta, z_{1,2}, z_{2,1})$ does not imply stability of system. Indeed, Figure 6 suggests that q_1 increases without bound, and by symmetry, so is q_2. Figure 7 shows that a substantial proportion of each pool has fluid

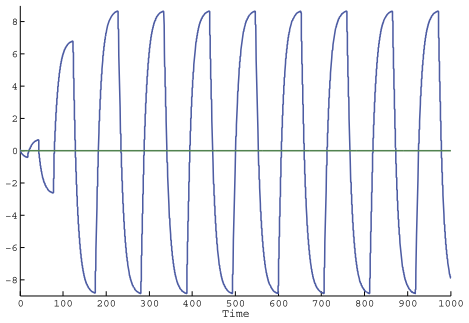

FIG 4. *increasing oscillations of the difference process* Δ *towards the periodic equilibrium, with* $\Delta(0) = 0.2 > \kappa$, $\theta = 0$ *and* $\mu = 0.1$.

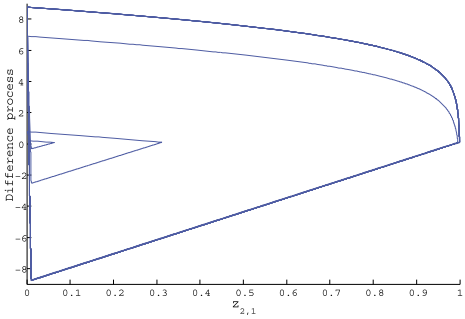

FIG 5. *spiraling of* $(z_{2,1}, \Delta)$ *outward towards the periodic equilibrium;* $\theta = 0$ *and* $\mu = 0.1$.

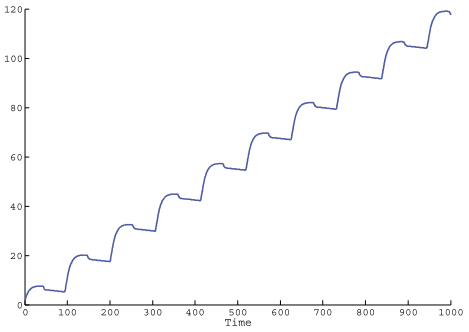

FIG 6. *trajectory of* q_1 *increases in oscillatory manner;* $\theta = 0$, $\mu = 0.1$.

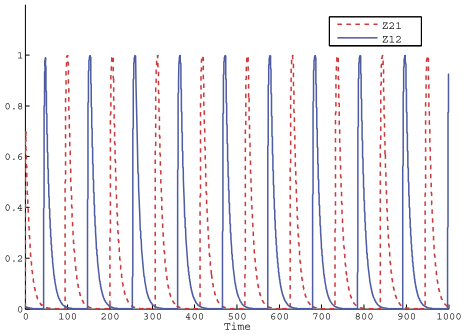

FIG 7. *sharing in both pools with* $z_{1,2}(0) = \tau = 0.01$ *and* $z_{2,1} = 0.005$, $\theta = 0$, $\mu = 0.1$.

TABLE 1
comparisons of the values obtained from the iterative algorithm for the approximating system in §6, to those of the iterative algorithm in §C.3.1 for the original system.

	$\Delta(0)$	$z(T_1)$	T_1	T_2
approximation	8.802	0.9992	7.093	46.044
original sys.	8.663	0.9992	7.270	46.044

from the other class for a non-negligible amount of time, which is the cause for the congestion collapse observed in Figure 6. See §F.

Finally, in Table 1 we compare the numerical solution to the iterative algorithm in §C.3.1 (in the "original sys." row), to the heuristic approximations developed in §6.4. We note that $L \approx 0.44 < \lambda = 0.98$ for L in (F.2).

7.2. *Bifurcation:* $\mu = 0.3$. The term "bifurcation" refers to a change in the equilibrium behavior of a dynamical system as the value of one of its parameters varies, while all other parameters remain unchanged. Following the analysis in §6, we now take the same system considered in §7.1 but change the value of μ. We do not carry out a full bifurcation analysis to find the bifurcation point in which the equilibrium behavior of the system changes, but instead consider a single value $\mu = 0.3$. To see how the system converges to the stationary point with no sharing, we change the initial condition in §7.1 and take $\Delta(0) = 20$. The trajectory of Δ is shown in Figure 8. (Note however, that we cut the vertical axis in this figure at the value 3 to make the oscillations more apparent.) Figure 9 shows the spiraling towards that equilibrium point in the $(z_{2,1}, \Delta)$ plane. Unlike the case depicted in Figure 5, now spiraling is "inward", i.e., the largest loop corresponds to the first cycle, and each of the four cycles is shorter than the previous one. we remark that the heuristic approximation in §6.4 was stopped in the fifth iterations since $\Delta^{(5)} < 0$.

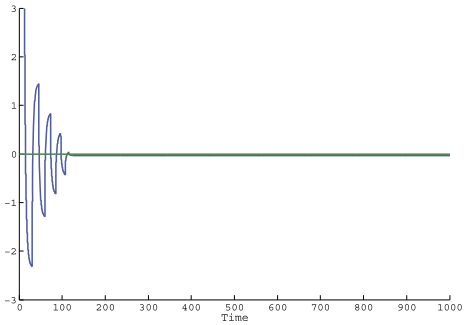

FIG 8. *decreasing oscillations of the difference process* Δ *towards its stationary point;* $\Delta(0) = 20$, $\mu = 0.3$, $\theta = 0$.

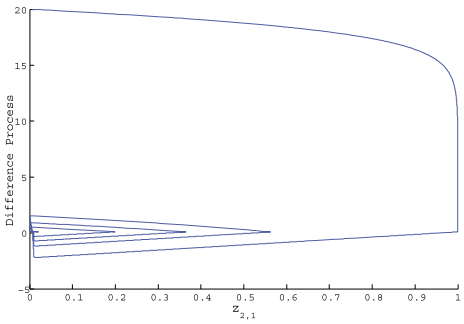

FIG 9. *spiraling "inward" of* $(z_{2,1}, \Delta)$ *to the stationary point,* $\mu = 0.3$, $\theta = 0$.

Observe that even though the convergence to the stationary point is fast in terms of the number of oscillations, it is very slow in continuous time. In particular, the system oscillates for more than a hundred time units before it ceases to oscillate.

7.3. *Adding abandonment.* For a numerical depiction of the approximating solution, we now consider a system with $\mu = 0.1$ as in §7.1 but add abandonment, taking $\theta = 0.01$. As can be seen by comparing Figures 10 and 11 to Figures 4 and 5, the system with no abandonment serves as a reasonable approximation for the a system with a small abandonment rate, but the oscillations are smaller, as is intuitively expected.

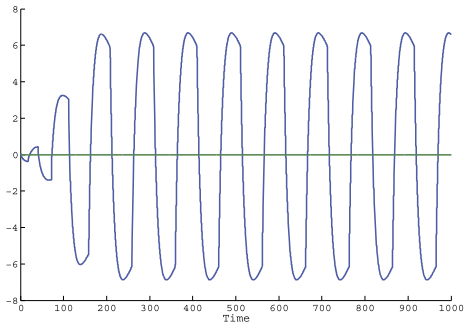

FIG 10. *Increasing oscillations of the difference process* Δ *towards the periodic equilibrium for the similar example as in Figure 4 but with positive abandonment rate* $\theta = 0.01$.

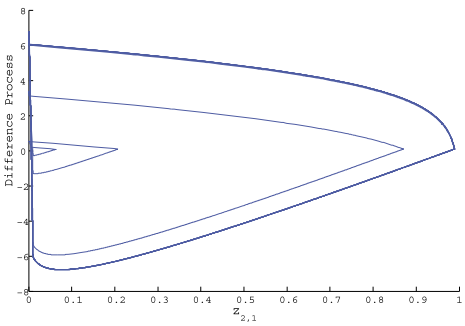

FIG 11. *image of* $(z_{2,1}, \Delta)$ *spiraling outward the periodic equilibrium for a similar example as in Figure 5, but with positive abandonment rate* $\theta = 0.01$.

7.4. *Simulations of systems with non-oscillating fluid limits.* So far we considered the fluid model (limit) alone. The numerical examples above show that congestion collapse can occur for very extreme parameter values μ and θ. In this section we show that the extreme examples provide important insights for cases for which the fluid limit never oscillates.

It is significant that for a given stochastic system X^n which is approximated by a fluid model x, there is freedom in how to choose the limiting thresholds. For example, if $n = 100$, then activation thresholds $k^n_{i,j} = 10$ can be considered as being $\sqrt{n}$ or as $0.1n$. In the latter case, the stochastic fluctuations are considered negligible with respect to the activation thresholds, and $\kappa = 0.1$. However, in the first case, $\kappa = 0$, and so the stochastic fluctuations are significant. Specifically, if $\kappa = 0$, then oscillations are much more likely to occur because $\mathbb{S}_{1,2} = \mathbb{S}_{2,1}$ in that case; see Remark 5.1.

System with a practically unstable stationary point. We simulated a system with similar parameters to those in §7.1 taking $n = 100$, so that there are 100 agents in each pool and $\lambda^n = 98$. As above, $\theta = 0.01$. Since $\kappa^n = 0.1n$, we take $\kappa^n = 10$, which we can also think of as being $\sqrt{n}$, i.e., $\kappa = 0$.

Figures 12 and 13 show a single sample path of the Q^n_1 process and the shared-customers processes for a system starting empty. Due to symmetry of the parameters and the initial condition of the two pools, the fluid model will unambiguously move through x^*_0. Once x^*_0 is hit, and since there is no sharing at that hitting time, the fluid model must remain at that point. However, random noise in the stochastic system causes sharing to begin, leading to extreme oscillations. From the fluid model perspective, this suggests that

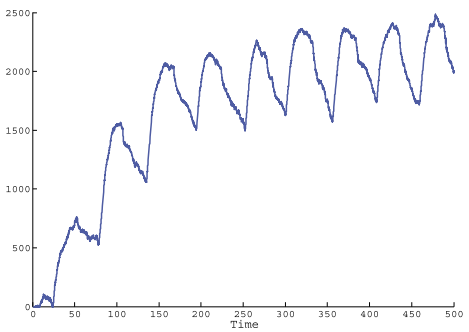

FIG 12. *simulated sample path of* Q_1^n*;* $n = 100$*,* $\lambda^n = 98$ $\kappa^n = 10$*,* $\theta = 0.01$*,* $\mu = 0.1$

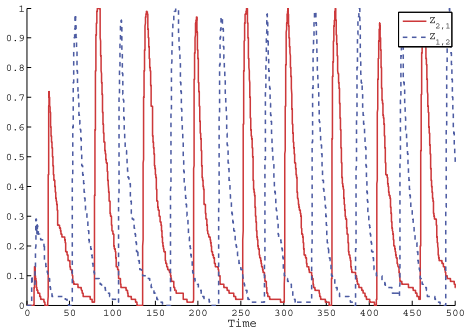

FIG 13. *simulated sample path of the number of shared customers in service in both pools;* $n = 100$*,* $\lambda^n = 98$ $\kappa^n = 10$*,* $\theta = 0.01$*,* $\mu = 0.1$

random fluctuations (that are negligible in fluid scale) quickly push the fluid limit from x_0^* to a state $\gamma \in \mathcal{O}$, leading to fluid-scaled fluctuations.

System with no oscillating solutions ($\mathcal{O} = \phi$). The fluid model gives important insight that cannot be obtained analytically even for systems with $\mathcal{O} = \phi$, i.e., systems that do not have oscillating fluid limits. We now take

$$n = 100: \ \lambda^n = 98, \ \mu = 0.5, \ \theta = 0.5, \ \tau^n = 1 \quad \text{and} \quad k_{i,j}^n = 10,$$

with the rest of the parameters being the same as in §7.1. The parameters θ and μ here are more likely in a practical call-center setting than the parameters in the examples above.

To show that $\mathcal{O} = \phi$ we solve the fluid model for an extreme example with $q_1(0) = 1$ and $q_2(0) = 1000$, $z_{1,2} = \tau$ and $z_{2,1} = 0$. In the simulation however, we have $Z_{2,1}^n = 20$ and $Z_{1,2}^n = 0$, which is a likely initial condition for a system recovering from an overload in queue 2. (The initial conditions of the stochastic system and the fluid model do not match because we want to show that the fluid model does not oscillate, and has no periodic equilibrium.)

Figure 14 shows a single sample path of the shared-customers processes from a single simulation run, and Figure 15 shows the fluid model of the system with the initial condition specified above. We only show figures of the shared customers service process, because both queues monotonically decrease to 0 in the fluid model, whereas customer abandonment make the oscillations of the queue processes unobservable in the simulation. From the practical point of view, this means that oscillations may be hard to detect in real time, unless one knows to look for them. Specifically, if the control parameters are chosen in accordance with the fluid model so as to ensure that

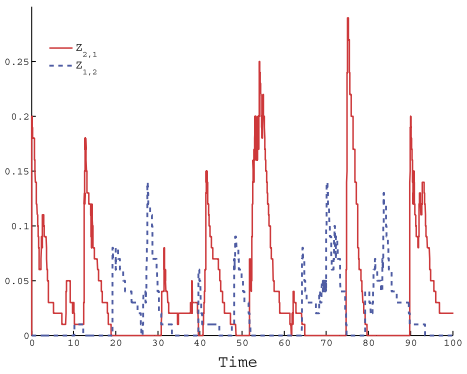

FIG 14. *simulated sample path of* $(Z^n_{1,2}, Z^n_{2,1})$ *when* $\mathcal{O} = \phi$*;* $n = 100$*,* $\lambda^n = 98$*,* $\mu = 0.5$*,* $\theta = 0.5$*,* $\tau^n = 1$ *and* $k^n_{i,j} = 10$

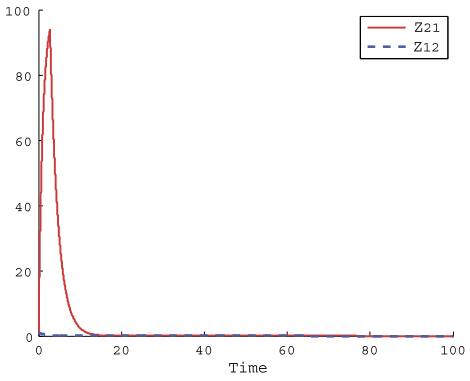

FIG 15. *fluid model* $(z_{1,2}, z_{2,1})$ *when* $\mathcal{O} = \phi$ *with an extreme initial condition is seen to converge to the stationary point with no sharing*

the system will not oscillate, and if only the queues are observed, then the oscillations may not be captured by the controller. Indeed, the controller may assume that sharing is initiated and stopped due to "legitimate" activations of the control to respond to changes in the arrival rates

We note that Figure 14 shows only the time interval $[0, 100]$ for clarity, but that the oscillations continued for the full run time of the simulation, which lasted 1500 time units. (As before, time here is measured in service time units $\mu_{i,i} = 1$, $i = 1, 2$.)

In ending we remark that the bad behavior shown here can be easily avoided by increasing $k^n_{i,j}$, as was discussed in Remark 5.1. A numerical example, related to the one given here, is given in Section 4.1 in [32]; see Figures 5 and 6 in that reference.

8. Implications of the fluid analysis for stochastic systems. In this section we consider the implications of our fluid analysis to the (finite) stochastic system which they approximate. These implications rely on the fact that the fluid model can be achieved as a *fluid limit* in the many-server heavy traffic limiting regime. That is, if $X^n(0) \Rightarrow x(0)$ in $\mathbb{R}_6$, then, uniformly on compact intervals, $\bar{X}^n \Rightarrow x$, where $\Rightarrow$ denotes convergence in distribution. See Theorem I.1 in §I.2 below for the precise statement and proof of this result.

Now, since for each fixed $n \geq 1$, X^n is clearly an irreducible and positive recurrent CTMC, it possesses a unique stationary distribution which is also its limiting distribution. In particular, for some random variable $X^n(\infty)$

with values in $\mathbb{R}_6$, it holds that, regardless of the initial condition,

$$X^n(t) \Rightarrow X^n(\infty) \quad \text{as} \quad t \to \infty.$$

Given that the fluid limit of $\bar{X}^n$ may oscillate indefinitely, and never converge to its stationary point, it is not a-priori clear whether $\bar{X}^n(\infty)$ can be approximated by x^* for large n. The following *weak law of large numbers* (WLLN) for the sequence $\{\bar{X}^n(\infty) : n \geq 1\}$, whose proof appears in §E, shows that the sequence of "fluid-scaled" stationary distributions converges to the stationary point x_0^* with no sharing, even if $\mathcal{O} \neq \phi$, i.e., the fluid limit may not converge to its stationary point x_0^*.

THEOREM 8.1 (WLLN for stationary distributions). $\bar{X}(\infty) \Rightarrow x_0^*$, *i.e., for each continuous and bounded function* $f : \mathbb{R}_6 \to \mathbb{R}$,

$$\lim_{n\to\infty} \lim_{t\to\infty} E[f(\bar{X}^n(t))] = f(x_0^*).$$

Note that taking the limits in Theorem 8.1 in the reverse order, namely, first taking $n \to \infty$ and then taking $t \to \infty$, is not possible when $\mathcal{O}$ is not empty, because the limit of $x(t)$ as $t \to \infty$ does not exist for all initial conditions. We therefore cannot prove Theorem 8.1 using standard arguments, as were laid out in the proof of Theorem 4 in [16].

8.1. *On the rate of convergence to stationarity.* The fact that $X^n \Rightarrow x$ uniformly on compact intervals together with Theorems 5.5 and 8.1 suggest that the state space of the *irreducible* CTMC X^n is nearly decomposable into two regions when $\mathcal{O} \neq \phi$. In particular, the chain may spend a long time in one region before eventually moving to the second region. For example, if $\bar{X}^n(0) \approx x(0) \in \mathcal{O}$ for n large, then X^n will approximately track the fluid trajectory with that initial condition. The oscillations of $\bar{X}^n$ can continue for arbitrarily large time periods as n increases.

On the other hand, if X^n is initialized with no sharing and no queues, then hitting the activation thresholds is a rare event asymptotically, and oscillations will not begin for a long time. However, the chain being irreducible, must eventually visit a state in an "oscillating region" for the CTMC, triggering oscillations that, as explained in the paragraph above, will take a long time before finally ending, if n is large.

To make this discussion rigorous, consider a sequence of initial conditions $\{X^n(0) : n \geq 1\}$ such that $\bar{X}^n(0) \Rightarrow x(0) \in \mathcal{O}$ as $n \to \infty$. Since $\bar{X}^n \Rightarrow x$ uniformly over compact intervals, and x is oscillating, we see that for any fixed $t > 0$ we can find N large enough, such that

$$\|\bar{X}^n(t) - \bar{X}^n(\infty)\|_{tv} > \epsilon, \quad \text{for all } n > N \text{ and for some } \epsilon > 0, \tag{8.1}$$

where $\|\cdot\|_{tv}$ denotes the total-variation norm (here given in terms of the random variables instead of their distributions); see, e.g., [10]. In particular, despite the fact that $\bar{X}^n(t) \Rightarrow \bar{X}^n(\infty)$ as $t \to \infty$ for any given n, and moreover, the convergence rate to stationarity is exponentially fast as we show below, the convergence rate to stationarity can be arbitrarily slow for a sufficiently large system.

To see that (8.1) indeed holds for all n large enough, note that convergence in total variation implies convergence in distribution (the two notions of convergence are in fact equivalent on countable state spaces). We can use the Lévy metric to measure distances between random variables corresponding to convergence in distribution. Specifically, we let the distance between two random variables X and Y with respective cumulative distribution functions F_X and F_Y, be

$$\begin{aligned} d_L(X,Y) &\equiv d_L(F_X,F_Y) \\ &\equiv \inf\{\epsilon > 0 : F_X(x-\epsilon) - \epsilon \le F_Y(x) \le F_X(x+\epsilon) + \epsilon \text{ for all } x\}. \end{aligned}$$

Then, for random variables Y and $\{Y^n : n \ge 1\}$, $Y^n \Rightarrow Y$ is equivalent to $d_L(Y^n, Y) \to 0$, and as mentioned above, if $\|Y^n - Y\|_{tv} \to 0$, then $d_L(Y^n, Y) \to 0$ as $n \to \infty$.

Now, take the contradictory assumption to (8.1), namely assume that there exists a time $t > 0$, such that

$$\|\bar{X}^n(t) - \bar{X}^n(\infty)\|_{tv} < \epsilon \quad \text{for all } n \ge 1 \text{ and } \epsilon > 0.$$

Then for this specific time t and for all n large enough, we have by the triangular inequality that

$$d_L(x(t), x_0^*) \le d_L(x(t), \bar{X}^n(t)) + d_L(\bar{X}^n(t), \bar{X}^n(\infty)) + d_L(\bar{X}^n(\infty), x^*) < 3\epsilon,$$

where the second inequality follows from Theorem I.1, our contradictory assumption and Theorem 8.1, and the above holds for any fluid trajectory, regardless of the initial condition. Hence, x_0^* is globally asymptotically stable, in contradiction to the assumption that $x(0) \in \mathcal{O}$.

The fact that X^n may converge extremely slowly to stationarity for large n is not entirely straightforward, because X^n is an exponentially ergodic CTMC, for each $n \ge 1$, and therefore considered to converge "fast". The proof of the following theorem can be found in §E.

THEOREM 8.2. *Fix $n \ge 1$. Then for any initial condition $k \in \mathbb{Z}_+^6$, there exist positive constants M_k and α (where M_k depends on the initial state k and α does not), such that*

$$\|X^n(t) - X^n(\infty)\|_{tv} \le M_k e^{-\alpha t} \tag{8.2}$$

9. Important takeaways for SSC-inducing controls. An SSC-inducing control for a stochastic system is commonly designed in order to achieve an optimal control for a diffusion approximation of the system. In particular, stochastic networks can rarely be analyzed exactly, even under a fixed control, and finding a "good" control that is nearly optimal (in an appropriate sense), and is furthermore implementable, is a prohibitively hard problem. The most prevalent approach to solving an optimal control problem, initiated by Harrison [17], is to solve a related *Diffusion Control Problem* (DCP) for the system in heavy traffic, and then translate the optimal control for the DCP into a control for the system it approximates. As reviewed in [7], a key step in this procedure is to formulate an *equivalent workload formulation* (EWF), in which the dimension of the workload process is reduced. Thus, an optimal control for the approximating diffusion process found via the DCP scheme, dictates an SSC-inducing control for the stochastic system, which forces the asymptotic workload process to a lower-dimensional "boundary" subset of its state space. In fact, the FQR-ART control considered in this paper was developed by an analogous approach: In [28] we optimized a fluid model of the stochastic system, and then proposed a control to achieve the optimal fluid solution.

Before describing general insights obtained from our analysis here, we mention that the FQR-ART control in the prelimit, and its limiting counterpart, share some of the general characteristics of SSC-inducing control and the resulting limiting control. First note that an optimal control for an EWF is necessarily "bang-bang" with a *singular* part in the sense of [19], namely, such a control uses the maximum "pushing" towards the boundary set when the queue is away from it, and switches instantaneously between the directions of pushing to maintain the queue on the boundary. Similarly, the control process in both FQR-ART and the resulting limiting control use the maximum force to push the queues toward the threshold and the boundary set, respectively. See also the paragraph below the display of $\mathbb{S}_{1,2}$ and $\mathbb{S}_{2,1}$ in §2.2, page 145.

We further observe that, in addition to the bang-bang portion, FQR-ART also has a "no-action" zone associated with states in which both difference processes are below the activation thresholds, or when one of the difference process is above its corresponding activation threshold, but the release threshold prevents the control from being activated. Our analysis here revealed that the no-action zone is responsible for the delayed activation of the control, and to the resulting oscillations when the control parameters are not chosen correctly. In that regard, we make the following general observation: While much attention had been given in the literature to the difficulties in

translating an optimal control for limiting approximations (typically diffusion approximations) to a control for the underlying sequence of stochastic systems, little attention was given to the difficulty in translating the asymptotic control parameters (when such a control is found) to a fixed system. Our analysis sheds light to possible problems that can arise in this second layer of translation. (Recall Remark 5.1 and the example in Figures 12–13. We discuss this issue further in §J in the appendix.)

We next demonstrate with specific examples how our insights apply to other models.

9.1. *A comparison with the N model.* To make the discussion concrete, we now compare some conceptual similarities between our model and the N model which was studied extensively in the single-server stations setting; see, e.g., [18, 20] and [2]. We refer to these references for other works on the N model in the conventional heavy traffic, and to [45] for a many-server heavy-traffic study of an N system. (We further remark that the N model is covered by the model considered in [15].)

The N model, depicted in Figure 3 above, has two *single-server* stations, with the server in station 1 dedicated to serving class-1 jobs, while the server in station 2 can handle both queues. (note that there are no arrows representing abandonment from the queues.) A (non-idling) service policy determines which queue server 2 should handle when both queues are non empty. Similarly to our simulation experiments in [28], which showed that the QIR control applied to a critically-loaded X system can lead to congestion collapse (see §2 above), a simulation in [18] demonstrated that a naive implementation of the $c\mu$ rule, under which server 2 always prioritizes the class 1 queue, may lead to congestion collapse in a critically-loaded N model.

A solution to a DCP for the N model led the authors in [2] to design a threshold policy and prove that this policy is asymptotically optimal and leads to SSC. Under that policy, server 2 helps queue 1 only when that queue is larger than a threshold, and otherwise this server only serves its own queue. It is further assumed that server 2 preempts the job it processes if it needs to switch the class it serves. Note that this control uses the maximum force to push queue 1 towards the threshold, and has a no-action zone associated with queue 1 being below that threshold. Since the threshold is taken to be $O(\log n)$ as $n \to \infty$, asymptotic SSC about the threshold implies that queue 1 is null in the diffusion limit. In particular, it is shown that the two-dimensional queue process has all its mass on the class-2 queue, and the idleness process has all its mass on server 2 (i.e., the class 1 queue is

asymptotically null, and server 1 never idle). Observe that the no-action zone shrinks to 0 as n increases, so that it does not appear in the limit.

It is significant that the asymptotics in [2] were achieved under the assumption that server 2 can switch in zero time between the two queues. This assumption is reasonable to make if the switching times in reality are sufficiently small relative to the service time, but it is intuitively clear that even short switching times will lead to diffusion-scaled chattering of queue 1 about the threshold. In fact, since server 2 switches infinitely-often between the two queues in the limit over any finite time interval, non-zero switching times may push the system to an overload, because switching times of server 2 essentially increase its idle period. Moreover, if the $O(\log n)$ threshold is not chosen appropriately for a given stochastic system, and in particular, if the the threshold is chosen to be too small, then, as in the example in [18], server 2 will again spend too much time helping queue 1, leading to congestion of its own queue.

Finally, the authors in [2] remark that a hysteresis control with two thresholds can be employed to avoid chattering of the queue about the threshold (see the remark at the bottom of p. 621 in this reference). Under this modified control, only once queue 1 crosses an upper threshold will server 2 switch to help this queue, and that help is turned off only once the queue returns to the lower threshold. It is again intuitively clear that, if the thresholds are not chosen appropriately for the fixed system, namely, if the lower threshold is too small and the upper one is too large, then server 2 may end up spending too much time helping queue 1, leading queue 2 to increase without bound in an oscillatory manner. Indeed, the two-thresholds control increases the "no-action" zone, in turn, leading to increased delays in activating the control.

10. Summary. In this paper we considered the FQR-ART overload control applied to the cyclic X model, when the control parameters are badly chosen. For the dynamical-system (fluid) limit, the purpose of the control is to attract any fluid trajectory to one of two sliding manifolds during overload periods, so as to maintain a pre-specified ratio between the two queues.

Switching fluid limit. We have shown that possible delays in activation and release of the control can lead to chattering and resulting oscillations, which translates to fluid-scaled fluctuations in the underlying stochastic system. The pathological oscillatory behavior can be analyzed via a switching dynamical system, as in Definition 3.1, within the framework of the many-server heavy-traffic FWLLN (Theorem I.1 in §I.2). Theorems 5.2 and 5.4, respectively, prove that the fluid limit has a unique stationary point and a

non-trivial periodic equilibrium that is associated with the oscillatory motion. Sufficient conditions for endless oscillations were provided in Theorem 5.5.

Fluid stability. In Theorem 5.3 it was shown that any fluid trajectory that ceases to oscillate must converge to the unique stationary point. A convenient approximating dynamical system to the fluid limit was developed and shown to be bi-stable in §6. Specifically, all the trajectories of the approximating system were shown to converge to one of the two equilibria – the stationary point x_0^* in (5.1), or a unique non-trivial periodic equilibrium. Finally, a simple heuristic construction in §6.4 can be used to approximate the values of the solutions to (3.10) at the switching times, and in particular, the values of the periodic equilibrium at the switching times, when it exists.

Implications. Numerical examples in §7 show the effectiveness of the approximating system. The simulation experiment in §7.4 demonstrates that our fluid model provides important insights into the untractable behavior of the underlying stochastic system, even when the fluid approximation itself is not oscillating.

From the theoretical stochastic perspective, the results in §8 demonstrate that, despite the fact that the stochastic system is an ergodic CTMC, and is even exponentially ergodic by Theorem 8.2, an oscillatory behavior of the fluid model implies that it may take very long time for the system to converge to stationarity. In particular, exponential ergodicity should does not necessarily imply "fast" convergence to stationarity.

From the practical perspective, the most important conclusion is that the control parameters must be chosen with caution. For example, the bad oscillatory behavior presented in §7.4 (which may be hard to detect in real time) can be avoided by choosing appropriate activation thresholds. We again refer to [32] for a further discussion.

APPENDIX A: OVERVIEW

This appendix contains supplementary material for the main paper. First, in §B we give notation for sets used in the paper, as well as sets that are used in the appendix. The proofs of the results in Section §5 appear in §C, and the proofs of the results in §6 are presented in §D. Both §§C and D include supporting results with their proofs, and efficient algorithms to compute the respective ODE's in switching times. The proofs of the theorems in §8 appear in §E. In §F we show how the approximating system can be employed to check whether congestion collapse occurs, and in §G we present an algorithm to compute the solution to the heuristic approximation suggested in §6.4. In §H we establish stronger forms of convergence of solutions to the

approximating system to their equilibrium behavior. In §I we show that the fluid model we considered in the main paper arises as the fluid limit in a many-server heavy-traffic fluid limit of the underlying model. The proof of the FWLLN is given in §I.2, after a brief expansion on the stochastic model and many-server scaling in §I.1. Finally, in §J we discuss implications of our results here for the control of the stochastic system.

APPENDIX B: NOTATION OF SETS

Below is a list of the different sets that appear in the paper. Their first appearance is in parenthesis.

- $\mathbb{S}^*$ – the set of all stationary points (§5.1).
- $\mathcal{M}$ – switching (or sliding) manifold in a general system (§1).
- $\mathcal{O}$ – the invariant set of oscillating solution, i.e., if $x(0) \in \mathcal{O}$, then x oscillates indefinitely (§5.1).
- $\mathcal{P}_{u^*}$ – the image of the periodic equilibrium u^* (§4).
- $\mathbb{S} \equiv [0, \lambda/\theta]^2 \times [0,1]^4$ – the state space of the fluid model (§2.2).
- $\mathbb{S}_{i,j}$ – the sliding manifold where $d_{i,j} = \kappa$ (§2.2)
- $\mathcal{S}_{u^*}$ – the stability region of the periodic equilibrium u^* (§4).
- $\mathcal{S}_{x^*}$ – the stability region of a stationary point x^* (§5.1).
- $\mathcal{S}_{x_0^*}$ – the stability region of the stationary point x_0^* in (5.1) (Theorem 5.2).
- $\mathbb{S}^a \equiv [0, \infty)^2 \times [0,1]^4$ – the state space of the approximating system (§6).
- $\mathbb{S}_\epsilon \equiv [\epsilon, \lambda/\theta]^2 \times [0, \tau]$, $\epsilon > 0$ – the state space of of solutions in $\mathcal{O}$ (§C.3.2).
- $\mathbb{S}_\kappa \equiv [\kappa + \epsilon_\kappa, \lambda/\theta] \times [0, \tau]$, where $\epsilon_\kappa > 0$ (Proof of existence part of Theorem 5.4 in §C.3.3).
- $\mathbb{S}_\mu \equiv [\Delta_\mu^M - \delta_\mu, \Delta_\mu^M]$, where Δ_μ^M is defined in (D.2) and δ_μ in (D.8) (Equation (D.5) in §D.2.3).

APPENDIX C: PROOF OF THE RESULTS IN SECTION 5

In this section we provide the proofs of the results in §5.

C.1. Proofs of Theorems 5.1, 5.2 and 5.3.

PROOF OF THEOREM 5.1. No sharing will ever occur because $q_i = (y_i - 1)^+$, and if $y_i(t) > 1$, so that the queue is positive, then $y_i(t)$ is decreasing at t, $i = 1, 2$. (Recall that $\lambda < \mu = 1$.) Hence, even if $d_{i,j}(0) = \kappa$ for $(i, j) = (1, 2)$ or $(i, j) = (2, 1)$, then $d_{i,j}(t) < \kappa$ for any $t > 0$ in some right-neighborhood of 0. It follows that, if $z_{i,j}(0) > 0$, $i \neq j$, then $z_{i,j}$ is strictly

decreasing, which implies that the service capacity in pool j is increasing. In turn, q_j must keep decreasing as long as it is strictly positive. Finally, since y_i is strictly decreasing as long as it is larger than λ and is strictly increasing otherwise, we have

$$y_i(t) \to \lambda \quad \text{as} \quad t \to \infty. \tag{C.1}$$

□

PROOF OF THEOREM 5.2. Suppose that $\gamma^* = (\gamma_i^*, \gamma_{i,j}^*; i,j = 1,2) \in \mathbb{S}^*$, is such that $\gamma^* \in \mathbb{S}_{1,2} \cup \mathbb{S}_{1,2}^+$, so that $\gamma_1^* \geq \kappa$. Consider the fluid model initialized at γ^*, i.e., $x(0) = \gamma^*$. If $z_{2,1}(0) = \gamma_{2,1}^* > 0$, then by the rules of FQR-ART, $\dot{z}_{2,1}(0) = -\mu_{2,1} z_{2,1}(0) < 0$, implying that $z_{2,1}$ is strictly decreasing. It follows that $\gamma_{2,1}^* = 0$, so that $\gamma_{1,1}^* = 1$ (because $\gamma_1^* \geq \kappa > 0$). But then

$$\dot{q}_1(0) = \lambda - \mu_{1,1}\gamma_{1,1}^* - \theta q_1(0) < \lambda - 1 < 0,$$

which contradicts the supposition that γ^* is a stationary point. Hence, $\mathbb{S}^* \cap (\mathbb{S}_{1,2} \cup \mathbb{S}_{1,2}^+) = \phi$. Similar arguments apply to $\mathbb{S}^* \cap (\mathbb{S}_{2,1} \cup \mathbb{S}_{2,1}^+)$. The same reasoning for $\gamma^* \in \mathbb{S}^* \cap \mathbb{S}_{1,2}^- \cap \mathbb{S}_{2,1}^-$ implies that $\gamma_{1,2}^* = \gamma_{2,1}^* = 0$ and $\gamma_1^* = \gamma_2^* = 0$. Then the arguments leading to (C.1) show that $\gamma^* = x_0^*$. Hence, we conclude that $\mathbb{S}^* = \{x_0^*\}$. □

PROOF OF THEOREM 5.3. Since $x(0) \in \mathcal{O}^c$ there exists a time $t_0 \geq 0$ such that $x(t) \notin \mathbb{S}_{1,2}^+ \cup \mathbb{S}_{2,1}^+$ for all $t \geq t_0$. If $x(t) \in \mathbb{S}_{i,j}^-$ for all $t \geq t_0$, then

$$\dot{z}_{i,j}(t) = -\mu z_{i,j}(t), \quad \text{so that } z_{i,j}(t) = z_{i,j}(t_0)e^{-\mu(t-t_0)}, \quad t \geq t_0.$$

Then both $z_{1,2}$ and $z_{2,1}$ converge to 0, and it is easy to see from (2.3) (recall that there is no new sharing taking place) that both queues will reach 0 in finite time. Then, after q_i reaches 0, all arriving fluid moves immediately into service, so that $\dot{z}_{2,2} = \lambda - z_{2,2}$, and we see that $z_{2,2}(t) \to \lambda$ as $t \to \infty$.

Now suppose that $x \in \mathbb{S}_{1,2}$ over an interval I. If $z_{2,1} > \tau$ over I, then no fluid flows from q_1 to pool 2, so that both queues evolve independently according to (2.3). Since $z_{1,2}$ and $z_{2,1}$ are strictly decreasing over I, the same arguments given above apply in this case. Therefore, assume that $z_{2,1} \leq \tau$ over an interval $J \subseteq I$ so that sharing is allowed. By Assumption 1, q_1 is strictly decreasing on J, and the sliding motion implies that $\dot{q}_1(t) - \dot{q}_2(t) = 0$, so that q_2 is strictly decreasing as well (at exactly the same rate as q_1). Now, some of the service capacity of pool 2 is given to queue-1 fluid at any point, so that, for $t \in J$,

$$\dot{q}_1(t) < \lambda - z_{1,1}(t) - \mu z_{2,1}(t) - \theta q_1(t)$$

$$\text{and} \quad \dot{q}_2(t) > \lambda - z_{2,2}(t) - \mu z_{1,2}(t) - \theta q_2(t).$$

Recalling that $q_1(t) = q_2(t) + \kappa$ and $z_{i,i}(t) = 1 - z_{j,i}(t)$ for $t \in J$, we have

$$0 = \dot{q}_1(t) - \dot{q}_2(t) < (1-\mu)(z_{2,1}(t) - z_{1,2}(t)) - \theta\kappa < (1-\mu)(z_{2,1}(t) - z_{1,2}(t)),$$

so that $z_{1,2}(t) < z_{2,1}(t)$. It follows that $z_{1,2}(t) \le \tau$ and is decreasing on J. In particular, both queues continue decreasing after the sliding motion is over.

The same arguments give that, if x ever slides on $\mathbb{S}_{2,1}$, then both queues are strictly increasing to 0. Hence, the processes $z_{1,2}$ and $z_{2,1}$ never increase above τ during sliding motion, so that both queues are strictly decreasing to 0. After q_i hits 0, $z_{j,i}$ decreases monotonically to 0 and $z_{i,i}$ converges to λ. □

C.2. Bounds to guarantee oscillations. We now provide auxiliary results, needed for the proof of Theorem 5.5, providing sufficient conditions for endless oscillations of solutions to (3.10) and congestion collapse. In §§C.2.1 and C.2.2 we construct simple bounds on T_1 and $x(T_1)$, and bounds on T_2 and the values of x over $[\Sigma_1, \Sigma_2)$, respectively. Universal bounds on the solution x and the holding times, and a numerical example, are given in §C.2.3.

C.2.1. *Auxiliary results: Bounds on T_1 and $x(T_1)$.* We can apply (3.16) to obtain bounds on T_1.

COROLLARY C.1 (bounds on T_1). *Under the initial conditions in Assumption 2, the interval end time T_1 is bounded above and below by*

$$0 < \frac{\theta\kappa + \Psi_L}{\theta\Delta(0) + \Psi_L} \le e^{-\theta T_1} \le \frac{\theta\kappa + \Psi_U}{\theta\Delta(0) + \Psi_U} < 1, \tag{C.2}$$

for Ψ_L and Ψ_U in (3.14), from which we deduce that

$$1 < \frac{\theta\Delta(0) + \Psi_U}{\theta\kappa + \Psi_U} \le e^{\theta T_1} \le \frac{\theta\Delta(0) + \Psi_L}{\theta\kappa + \Psi_L} < \infty,$$

and

$$0 < \log\left(\frac{\theta\Delta(0) + \Psi_U}{\theta\kappa + \Psi_U}\right) \le \theta T_1 \le \log\left(\frac{\theta\Delta(0) + \Psi_L}{\theta\kappa + \Psi_L}\right) < \infty.$$

The associated bounds on T_1, denoted by $T_1^L \equiv T_1^L(\Delta(0))$ and $T_1^U \equiv T_1^U(\Delta(0))$, are both strictly increasing functions of $\Delta(0)$, both approaching 0 as $\Delta(0) \downarrow \kappa$ and ∞ as $\Delta(0) \uparrow \infty$. In particular,

$$T_1^L \equiv \left(\frac{1}{\theta}\right)\log\left(\frac{\theta\Delta(0) + \Psi_U}{\theta\kappa + \Psi_U}\right) = \left(\frac{1}{\theta}\right)\log\left(1 + \frac{\Delta(0) - \kappa}{(\Psi_U/\theta) + \kappa}\right) \le \frac{\Delta(0) - \kappa}{\Psi_U + \theta\kappa}$$

and

$$T_1^U \equiv \left(\frac{1}{\theta}\right)\log\left(\frac{\theta\Delta(0)+\Psi_L}{\theta\kappa+\Psi_L}\right) = \left(\frac{1}{\theta}\right)\log\left(1+\frac{\Delta(0)-\kappa}{(\Psi_L/\theta)+\kappa}\right) \le \frac{\Delta(0)-\kappa}{\Psi_L+\theta\kappa}$$

so that

$$\begin{aligned} 0 < T_1^U - T_1^L &= \left(\frac{1}{\theta}\right)\left(\log\left(1+\frac{\Delta(0)-\kappa}{(\Psi_L/\theta)+\kappa}\right) - \log\left(1+\frac{\Delta(0)-\kappa}{(\Psi_U/\theta)+\kappa}\right)\right) \\ &= \left(\frac{1}{\theta}\right)\left(\log\left(\frac{\theta\Delta(0)+\Psi_L}{\theta\kappa+\Psi_L}\right)\left(\frac{\theta\kappa+\Psi_U}{\theta\Delta(0)+\Psi_U}\right)\right). \end{aligned}$$

PROOF. Exploit (3.16) with the equation $\Delta(T_1) = \kappa$ characterizing T_1. □

The bounds we have just obtained on T_1 can be used to obtain bounds on $q_1(T_1)$. Recall that $\kappa < \Delta(0)$ and $\Psi_L < \Psi_U < 0$. Applying (C.2) with (3.6), we immediately obtain

COROLLARY C.2 (bounds on $q_1(T_1)$). *$q_1(t)$ is bounded from below by q_1^L and from above by q_1^U, where, for Ψ_L and Ψ_U in (3.14),*

$$\begin{aligned} 0 < q_1^L(T_1) &\equiv \frac{\lambda}{\theta} - \left(\frac{\lambda}{\theta} - q_1(0)\right)\left(\frac{\theta\kappa+\Psi_L}{\theta\Delta(0)+\Psi_L}\right) \\ &\le q_1(T_1) \le \frac{\lambda}{\theta} - \left(\frac{\lambda}{\theta} - q_1(0)\right)\left(\frac{\theta\kappa+\Psi_U}{\theta\Delta(0)+\Psi_U}\right) \equiv q_1^U(T_1) < \infty. \end{aligned}$$

Similarly, Applying (3.4), we have

COROLLARY C.3 (bounds on $z_{2,1}(T_1)$).

$$0 < z_{2,1}^L(T_1)) \equiv 1 - e^{-T_1} < z_{2,1}(T_1)) < 1 - (1-\tau)e^{-T_1} \equiv z_{2,1}^U(T_1)) < 1.$$

C.2.2. *Bounds on T_2 and $\{x(t) : T_1 \le t \le T_1 + T_2)$.* For bad oscillatory behavior, we will want to see that $q_2(T_1 + t)$ remains positive and, furthermore that $d_{2,1} < 0$. to ensure that the initial conditions in Assumption 2 hold at the switching time $\Sigma_2 \equiv T_1 + T_2$ with the index labels reversed. From Corollary C.2, we obtain the following

COROLLARY C.4 (lower bounds on the queue lengths on $[T_1, T_1 + T_2)$).

$$q_2(T_1) - \kappa = q_1(T_1) \ge q_1^L(T_1) = \frac{\lambda}{\theta} - \left(\frac{\lambda}{\theta} - q_1(0)\right)\left(\frac{\theta\kappa+2}{\theta\Delta(0)+2}\right),$$

so that, for $i = 1, 2$,

$$q_i(T_1 + t) \geq q_1^L(T_1)e^{-\theta t} - \left(\frac{1-\lambda}{\theta}\right)(1 - e^{-\theta t})$$
$$= \left(\frac{\lambda}{\theta} - \left(\frac{\lambda}{\theta} - q_1(0)\right)\left(\frac{\kappa + 2}{\Delta(0) + 2}\right)\right)e^{-\theta t} - \left(\frac{1-\lambda}{\theta}\right)(1 - e^{-\theta t}),$$

which is a strictly decreasing function of t. *As a consequence, a sufficient condition for both* $q_1(t)$ *and* $q_2(t)$ *to remain positive throughout* $[T_1, T_1 + T_2]$ *is for*

$$\left(\frac{\lambda}{\theta} - \left(\frac{\lambda}{\theta} - q_1(0)\right)\left(\frac{\theta\kappa + 2}{\theta\Delta(0) + 2}\right)\right)e^{-\theta T_2} > \left(\frac{1-\lambda}{\theta}\right)(1 - e^{-\theta T_2}),$$

for which a sufficient condition is

$$\left(\frac{\lambda}{\theta} - \left(\frac{\lambda}{\theta} - q_1(0)\right)\left(\frac{\theta\kappa + 2}{\theta\Delta(0) + 2}\right)\right)e^{-\theta T_2^U} > \left(\frac{1-\lambda}{\theta}\right)(1 - e^{-\theta T_2^U}),$$

where

$$T_2^U \equiv \frac{\log_e\left([1 - (1 - z_{2,1}(0))e^{-T_1^U}]/\tau\right)}{\mu} \leq \frac{\log_e\left([1 - (1-\tau)e^{-T_1^U}]/\tau\right)}{\mu}.$$

for T_1^U *in Corollary* C.1.

C.2.3. *Universal bounds.* We now consider the performance over a range of initial conditions. First, we introduce lower and upper bounds on the initial difference $\Delta(0) \equiv q_2(0) - q_1(0)$. We assume that

$$0 < \kappa < \Delta_L(0) \leq \Delta(0) \leq \Delta_U(0) < \infty \tag{C.3}$$

uniformly enforcing Assumption 2. We also assume that the smaller queue length is bounded below and above by

$$0 < q_1^L(0) \leq q_1(0) \leq q_1^U(0) < \frac{\lambda}{\theta} < \infty, \tag{C.4}$$

again uniformly enforcing Assumption 2.

Now let T_1^{L*} be the lower bound T_1^L for T_1 in Corollary C.1 when $\Delta(0) = \Delta_L(0)$ and let T_1^{U*} be the lower bound T_1^U for T_1 in Corollary C.1 when $\Delta(0) = \Delta_U(0)$.

LEMMA C.1 (universal bounds on T_1). *For all initial conditions satisfying* (C.3) *and* (C.4),

$$0 < T_1^{L*} \leq T_1 \leq T_1^{U*} < \infty.$$

PROOF. Apply Corollary C.1. □

LEMMA C.2 (universal bounds on $z_{2,1}(T_1)$ and T_2). *If, together with* (C.3) *and* (C.4),

$$1 - e^{-T_1^{L*}} > \tau, \tag{C.5}$$

then

$$1 - e^{-T_1} > \tau, \quad \tau < z_{2,1}(T_1^{L*}) \le z_{2,1}(T_1) \le z_{2,1}(T_1^{U*})$$

and

$$T_2^{L*} \equiv \frac{\log_e\left(z_{2,1}(T_1^{L*})/\tau\right)}{\mu} \le T_2 \le \frac{\log_e\left(z_{2,1}(T_1^{U*})/\tau\right)}{\mu} \equiv T_2^{U*} \tag{C.6}$$

for all initial conditions satisfying (C.3) *and* (C.4).

PROOF. Apply (3.4) and (3.21) together with Lemma C.2. □

If a periodic equilibrium exists, then the value of $z_{1,2}(\Sigma_2)$ will equal to $z_{2,1}(\sigma_2)$ on that equilibrium, as explained below (3.2) in §3. See also (5.2) in Theorem 5.4. We put the results above together to obtain bounds on $z_{1,2}(T_1 + T_2)$, which will serve as the new value of $z_{2,1}(0)$ in a continuation of the algorithm beyond time $\Sigma_2 = T_1 + T_2$.

LEMMA C.3 (universal bounds on $z_{1,2}$ (Σ_2)). *If conditions* (C.3), (C.4) *and* (C.5) *hold, then*

$$\begin{aligned} 0 < z_{1,2}^{L*}(T_1 + T_2) &\equiv e^{-\mu T_1^{U*}} z_{2,1}(T_1^{U*}) \le z_{1,2}(T_1 + T_2) \le e^{-\mu T_1^{L*}} z_{2,1}(T_1^{L*}) \\ &\equiv z_{1,2}^{U*}(T_1 + T_2) < \tau \end{aligned}$$

for all initial conditions satisfying (C.3) *and* (C.4).

PROOF. Apply (3.22) together with the lemmas above. □

Next we consider the queue lengths at time $T_1 + T_2$.

LEMMA C.4 (universal lower bounds on the queue lengths at time T_1+T_2). *If* (C.3), (C.4) *and* (C.5) *hold, then*

$$q_2(T_1) - \kappa = q_1(T_1) \ge q_1^{L*}(T_1) \equiv \frac{\lambda}{\theta} - \left(\frac{\lambda}{\theta} - q_1^L(0)\right)\left(\frac{\theta\kappa + 2}{\theta\Delta^L(0) + 2}\right),$$

for all initial conditions satisfying (C.3) *and* (C.4), *where* $q_1^L(0)$ *and* $\Delta^L(0)$ *are given in* (C.3) *and* (C.4). *If, in addition,*

$$q_1^{L*}(T_1+T_2) \equiv q_1^{L*}(T_1)e^{-\theta T_2^{U*}} > \left(\frac{1-\lambda}{\theta}\right)(1-e^{-\theta T_2^{U*}}), \tag{C.7}$$

then the two queue lengths $q_1(t)$ *and* $q_2(t)$ *remain positive throughout* $[T_1, T_1+T_2]$ *for all initial conditions satisfying* (C.3) *and* (C.4).

PROOF. Apply Corollary C.4 and (C.6). □

Finally, we obtain lower and upper bounds on the queue difference at time T_1+T_2.

LEMMA C.5 (universal bounds on the queue difference at time T_1+T_2). *If conditions* (C.3), (C.4) *and* (C.5) *hold, then*

$$\begin{aligned}\Delta_L(T_1+T_2) &\equiv \kappa e^{-\theta T_2^{U*}} - A_U\left(\frac{e^{-\theta T_2^{L*}} - e^{-\mu T_2^{U*}}}{\mu-\theta}\right)\\ &\le \Delta(T_1+T_2) \le \Delta_U(T_1+T_2)\\ &\equiv \kappa e^{-\theta T_2^{L*}} - A_L\left(\frac{e^{-\theta T_2^{U*}} - e^{-\mu T_2^{LU*}}}{\mu-\theta}\right)\end{aligned}$$

for all initial conditions satisfying (C.3) *and* (C.4), *where* T_2^{L*} *and* T_2^{U*} *are given in* (C.6) *and*

$$A_L \equiv (1-\mu)(z_{1,2}^{L*}(T_1) - z_{2,1}^{U*}(T_1)) \le A \le (1-\mu)(z_{1,2}^{U*}(T_1) - z_{2,1}^{L*}(T_1)) \equiv A_U$$

for A *in* (3.18).

A numerical example. Consider the bounds in Lemma C.5. Since κ is taken to be relatively small,

$$\Delta_L(T_1+T_2) \approx A_U\left(\frac{e^{-\theta T_2^{L*}} - e^{-\mu T_2^{U*}}}{\mu-\theta}\right).$$

In addition, $A_U \le (1-\mu)(\tau-1)$, so that, for given μ and τ, A in this lemma is bounded from above by a constant. These observations help to determine an initial value $\Delta_L(0)$ for which (C.14) will be satisfied. For the same parameters in §7 $\mu = 0.1$, $\lambda = 0.98$, $\tau = 0.01$, $\kappa = 0.1$ and $\theta = 0.01$, the constant bound of A_U is -0.891 and $\Delta_L(T_1+T_2) \ge 6.21$. Hence, (C.14) holds for some values of $\Delta(0)$ in the interval $(\kappa, 6.21)$. For example, taking $\Delta_L(0) = 4$, $\Delta_U(0) = 7$ and $q_1^L(0) = 1$, we obtain $\Delta_L(\Sigma_1) \approx 6 > \Delta_L(0)$ and $q_1^L(\Sigma_1) = 1.8 > q_1^L(0)$.

C.3. Proofs of Theorems 5.4 and 5.5. To establish these results, we exploit an algorithm for efficiently computing a solution to the switching model in (3.10) and efficiently calculating the periodic equilibrium if it exists. The algorithm improves on the piecewise numerical solution of the piecewise ODE in (3.10) by exploiting the exact formulas in §3. We can recursively calculate the values at the switching times Σ_i and then afterwards calculate the trajectory in between. By iterating, we can easily determine numerically if the solution converges to the stationary point or not. Numerical experience indicates that if the solution oscillates indefinitely, then it rapidly converges to a periodic equilibrium. In particular, the algorithm identifies the periodic equilibrium. However, more is required to provide a mathematical proof of existence, uniqueness and convergence.

C.3.1. *An efficient algorithm for the periodic equilibrium.* A periodic equilibrium u^* has an important closure property: If $u^*(t)$ satisfies Assumption 2 for some t, then $u^*(t + \Sigma_4) = u^*(t)$. Due to the symmetry of our model, we can relate the system state at time $t + \Sigma_2$ to the system state at time t. The state at time $t + \Sigma_2$ should coincide with the state at time t with the labels reversed. That is, we should have

$$\begin{aligned} &q_1(t + \Sigma_2) = q_2(t) > 0, \quad q_2(t + \Sigma_2) = q_1(t) > 0 \\ &z_{1,2}(t + \Sigma_2) = z_{2,1}(t) \quad \text{and} \quad z_{2,1}(t + \Sigma_2) = z_{1,2}(t) = \tau. \end{aligned} \tag{C.8}$$

with the condition that the pools remain full throughout:

$$z_{1,1}(s) + z_{2,1}(s) = 1 \quad \text{and} \quad z_{2,2}(s) + z_{1,2}(s) = 1, \quad 0 \le s \le t + \Sigma_2.$$

(Observe that the labels of the processes in the second equality in (5.2) are reversed.) If indeed we can establish the closure property in (C.8), then we will have proved that there exists a periodic equilibrium.

It is natural to search for the equilibrium by iterating: We pick a candidate initial vector $x_3(0) \equiv (q_1(0), q_2(0), z_{2,1}(0))$, letting $z_{1,2}(0) = \tau$, so that Assumption 2 holds. We then solve for T_1, T_2, and $(q_1(T_1 + T_2), q_2(T_1 + T_2), z_{1,2}(T_1 + T_2))$, as indicated above. we then redefine $(q_1(0), q_2(0), z_{2,1}(0))$ to be $(q_2(T_1 + T_2), q_1(T_1 + T_2), z_{1,2}(T_1 + T_2))$ and repeat the calculation.

If at some iteration we obtain an unreasonable value for x_3, e.g., $q_i < 0$, $i = 1$ or $i = 2$, or $\Delta \le \kappa$, then the algorithm is stopped and we conclude that the solution corresponding to the initial condition we chose converges to x_0^* (due to Theorem 5.3). However, a pathological case has $\Delta > \kappa$ for all iterations, but $\Delta \to \kappa$. Let Δ^* and T_1^* denote the limit of Δ and T_1 when the algorithm is iterated indefinitely. Observe that $\Delta^* = \kappa$ implies $T_1^* = 0$, so that the corresponding limiting solution u^* is necessarily a constant function.

This case is clearly a pathology, due to the uniqueness of the stationary point x_0^*. The following lemma ensures that such a pathological behavior of the algorithm is not possible. In particular, if at some iteration of the algorithm Δ is too close to κ, then this is also the last iteration.

LEMMA C.6. *There exists $\epsilon_\kappa > 0$ such that, if $\kappa < \Delta(0) < \kappa + \epsilon_\kappa$, then $x(\Sigma_2) > -\kappa$. In particular $x(0) \in \mathcal{O}^c$, so that $x(t) \to x_0^*$ as $t \to \infty$.*

PROOF. By Lemma 3.1, Δ is bounded from above by the linear function $-\Psi_L$. Hence, for any $\delta_1 > 0$ we can find $\epsilon_1 > 0$ such that, if $\kappa < \Delta(0) < \kappa+\epsilon_1$, then $0 < T_1 < \delta_1$. The explicit expressions of $z_{2,1}$ in (3.4) and T_2 in (3.21) show that, for any $z_{2,1}(0)$ and $\delta_2 > 0$, we can choose δ_1 sufficiently small to ensure that $T_2 < \delta_2$ (even if $T_2 > 0$). Hence, for any $\delta > 0$, we can find $\epsilon > 0$ such that, if $\kappa < \Delta(0) < \kappa + \epsilon$, then $\Sigma_2 < \delta$, by first choosing δ_2 and then an appropriate δ_1 to ensure that $\delta_1 + \delta_1 \leq \delta$. The continuity of Δ implies that there exists a $\delta_\kappa > 0$ such that, if $\Sigma_2 < \delta_\kappa$, then $\Delta(\Sigma_2) > -\kappa$. It follows that for all t in some right neighborhood of Σ_2 both $z_{1,2}(t)$ and $z_{2,1}(t)$ are strictly less than τ, so that both queues are strictly decreasing.

Now, if x ever hits $\mathbb{S}_{i,j}$, $(i,j) = (1,2)$ or $(i,j) = (2,1)$, after time Σ_2, then it can not cross it to $\mathbb{S}_{i,j}^+$. To see this, suppose for example that x hits $\mathbb{S}_{2,1}$ at some time $t > \Sigma_2$. Since x evolves according to the ODE's (3.4) - (3.5) when in $\mathbb{S}_{2,1}^+$, the derivative of $\Delta(t) \in \mathbb{S}_{2,1}^+$ is strictly negative; see Lemma 3.1. Moreover, sharing is allowed to start immediately because $z_{1,2} < \tau$. Therefore, if $\Delta(0) < \kappa + \epsilon_\kappa$, then $x(0) \in \mathcal{O}^c$, so that $x(t) \to x_0^*$ as $t \to \infty$ by Theorem 5.3. □

Let $\Delta^{(k)}$ be the value of Δ at the k^{th} iteration of the algorithm. It follows from Lemma C.6 that

COROLLARY C.5. *If $x(0) \in \mathcal{O}$, then $\Delta^{(k)} \in [\kappa + \epsilon_\kappa, \lambda/\theta]$, $k \geq 1$, for $\epsilon_\kappa > 0$ in Lemma C.6.*

C.3.2. *Proof of Theorem 5.5.*

PROOF. We first impose conditions on the model parameters and initial conditions so that the iterative algorithm in §C.3.1 mapping the initial state vector $x_3(0) \equiv (q_1(0), q_2(0), z_{2,1}(0))$ into the state vector $x_3(\Sigma_2) \equiv (q_1(\Sigma_2), q_2(\Sigma_2), z_{1,2}(\Sigma_2))$ and then iterated again to map $x_3(0)$ into $x_3(\Sigma_4) \equiv (q_1(\Sigma_4), q_2(\Sigma_4), z_{2,1}(\Sigma_4))$ is a map of the convex compact subset $\mathbb{S}_\epsilon$ of the Euclidean space $\mathbb{R}_3$ into itself, where $\mathbb{S}_\epsilon$ is the subset $\mathbb{S}_\epsilon \equiv [\epsilon, \lambda/\theta] \times [\epsilon, \lambda/\theta] \times [0, \tau]$ for some $\epsilon > 0$.

For that purpose, we introduce lower and upper bounds on the initial queue difference $\Delta(0)$,

$$0 < \kappa < \Delta_L(0) \le \Delta(0) \equiv q_2(0) - q_1(0) \le \Delta_U(0) < \infty, \tag{C.9}$$

and assume that the smaller queue length $q_1(0)$ is bounded below as well as above by

$$0 < q_1^L(0) \le q_1(0) \le q_1^U(0) < \frac{\lambda}{\theta} < \infty, \tag{C.10}$$

both consistent with Assumption 2.

We can apply (3.16) to establish upper and lower bounds on T_1, as shown in Corollary C.1. Those bounds are

$$T_1^L \equiv \left(\frac{1}{\theta}\right) \log\left(\frac{\theta\Delta_L(0) + \Psi_U}{\theta\kappa + \Psi_U}\right) \le T_1 \le T_1^U \equiv \left(\frac{1}{\theta}\right) \log\left(\frac{\theta\Delta_U(0) + \Psi_L}{\theta\kappa + \Psi_L}\right) \tag{C.11}$$

where $\Delta_L(0)$ and $\Delta_U(0)$ come from (C.9) and Ψ_U and Ψ_L are upper bounds on Ψ in (3.13) and (3.14). We then impose an upper bound on τ by requiring $\tau < 1 - e^{-T_1^L}$, which imposes an upper bound on T_2, i.e.,

$$T_2 \le T_2^U \equiv \frac{\log_e\left(z_{2,1}(T_1^U)/\tau\right)}{\mu}. \tag{C.12}$$

If, in addition,

$$\begin{aligned} q_1^L(T_1 + T_2) &\equiv \left[\frac{\lambda}{\theta} - \left(\frac{\lambda}{\theta} - q_1^L(0)\right)\left(\frac{\theta\kappa + 2}{\theta\Delta^L(0) + 2}\right)\right] e^{-\theta T_2^U} \\ &> \left(\frac{1-\lambda}{\theta}\right)(1 - e^{-\theta T_2^U}), \end{aligned} \tag{C.13}$$

then the two queue lengths both remain positive throughout the interval $[0, T_1 + T_2]$ and $q_1(T_1 + T_2) \ge q_1^L(T_1 + T_2)$ in (C.13), as shown in Lemma C.5. (If necessary, we redfine $q_1^L(0)$ so that $q_1^L(T_1 + T_2) \ge q_1^L$ as well as (C.10).) Finally, if

$$0 < \kappa < \Delta_L(0) \le \Delta(T_1 + T_2) \equiv q_2(0) - q_1(0) \le \Delta_U(0) < \infty, \tag{C.14}$$

then we can iterate without limit, with $\Sigma_q = \infty$. Condition (C.14) can be checked after the first iteration. However, sufficient conditions for (C.14) to hold without performing the first iteration are given in Lemma C.5. Numerical examples confirm that all these conditions can be satisfied, thus proving Theorem 5.5. □

C.3.3. *Proof of Theorem 5.4.*

PROOF. For a solution x with $x(0) \in \mathcal{O}$, $\Sigma_q = \infty$, so that the algorithm can be iterated indefinitely. In each iteration, the algorithm acts as a map of the vector $x_3(0) = (q_1(0), q_2(0), z_{2,1}(0))$ to $x_3(\Sigma_4)$ (with $x_3(\Sigma_4)$ serving as the initial condition for the following iteration). Therefore, the algorithm maps the compact and convex set $[0, \lambda/\theta] \times [\kappa, \lambda/\theta] \times [0, \tau]$ into itself. As long as the solution oscillates, we can restrict attention to the two-dimensional process $x_2 \equiv (\Delta, z_{2,1})$, because $\Delta(0) = \Delta(\Sigma_4) = \kappa$. In particular, at each iteration of the algorithm we compute $\Delta(\Sigma_2)$ and use it as the initial condition for the next iteration.

Corollary C.5 implies that for this two-dimensional process x_2, the algorithm acts as a map from the space $\mathbb{S}_\kappa \equiv [\kappa + \epsilon_\kappa, \lambda/\theta] \times [0, \tau]$ into itself, where $\epsilon_\kappa > 0$. The explicit solution to the ODE (3.10) over $[0, \Sigma_4]$ and to Δ in (3.15) and (3.19) shows that this map is continuous. Hence, by Brouwer's fixed point theorem (e.g., Theorem 5.28 in [36]) there exists a fixed point to this map in the set $\mathbb{S}_\kappa$. That fixed point cannot be also a fixed point of (3.10), due to Theorem 5.2, i.e., due to the uniqueness of x_0^*. It follows that there exists a solution to (3.10) satisfying (C.8) which is not a constant. Necessarily, such a solution is a non-trivial periodic equilibrium. □

APPENDIX D: PROOF OF THE RESULTS IN SECTION 6

D.1. Proof of Lemma 6.1. Define the function $F : B \to \mathbb{R}_+$, where

$$B \equiv (\kappa, \infty) \times (0, \infty) \quad \text{and} \quad F(\Delta, T) \equiv \frac{\Delta - 1 + \mu - \kappa}{1 + \mu} + \frac{1 - \mu}{1 + \mu} e^{-T} - T,$$

and the function

$$h(T) \equiv \Delta - 1 + \mu - \kappa + (1 - \mu)e^{-T} - (1 + \mu)T.$$

Note that $h(0) > 0$ and $h(T) \to -\infty$ as $T \to +\infty$. Furthermore, $h'(T) < 0$, so that $h(T)$ is strictly decreasing.

It follows that for any fixed $\Delta > \kappa$, there exists a unique $T > 0$, such that $(\Delta, T) \in B$ and $F(\Delta, T) = 0$. In addition, it clearly holds that $\frac{\partial F}{\partial \Delta}$ and $\frac{\partial F}{\partial T}$ exist in B and are continuous, and that $\frac{\partial F}{\partial T} \neq 0$ for all real T. Then by the implicit-function theorem there exists a unique continuously-differentiable function $T(\Delta)$, such that $F(\Delta, T(\Delta)) = 0$ over the domain B, and

$$\frac{dT}{d\Delta} = -\frac{\frac{\partial}{\partial \Delta} F}{\frac{\partial}{\partial T} F} = \frac{1}{(1 - \mu)e^{-T} + (1 + \mu)} > 0,$$

so that T is strictly increasing in Δ.

In passing we note that the point $(\Delta_0, T_0) \equiv (1 - \mu + \kappa, 0)$ satisfies $F(\Delta_0, T_0) = 0$. However, this point is not in B, so there is no contradiction to the claim that there exists a function $T(\Delta)$ as in the proof of Lemma 6.1. □

D.2. Proof of Theorems 6.1 and 6.2.

D.2.1. *Proof of Theorem 6.1.* Recall that the ODE (6.17) is solved until time Σ_4^a, and can then be continued beyond that time by taking $x_3^a(0) \equiv x_3^a(\Sigma_4^a)$ to be a new initial condition provided that $x_3^a(\Sigma_4^a)$ satisfies (6.3), i.e. if $\Delta^a(\Sigma_4^a) > \kappa$. However, if $\Delta^a(\Sigma_4^a) \leq \kappa$, then the ODE does not follows the switching pattern in (6.17). The next lemma shows that, in this case, the solution will converge to x_0^* and will therefore cease to oscillate.

LEMMA D.1. *If $\Delta^a(0) \leq \kappa$, but all other conditions in (6.3) hold, then $x^a(t) \to x_0^*$ for x_0^* in (5.1).*

Note that the lemma considers the full six-dimensional approximation x^a, and not only the three-dimensional restriction x_3^a.

PROOF. The initial condition has $z_{1,2}^a(0) = z_{2,1}^a(0) = 0$, so that $z_{1,1}^a(0) = z_{2,2}^a(0) = 1$. Hence, both pools serve only their own fluid queues, as long as $q_i(t) - q_j(t) < \kappa$, for both $(i,j) = (1,2)$ and $(i,j) = (2,1)$. Therefore (see (2.3))

$$\dot{q}_1(t) = \dot{q}_2(t) = \lambda - 1 < 0, \quad 0 \leq t < \Sigma_q^a,$$

so that $\dot{\Delta}^a(t) = 0$ on $[0, \Sigma_q^a)$, and no sharing can begin during that interval. At time Σ_q^a at least one of the queues hits 0, say q_i^a. If the other queue is still positive at that time, then it continues to decrease at the same constant rate as before. Since $|q_i^a(\Sigma_q^a) - q_j^a(\Sigma_q^a)| = q_j^a(\Sigma_q^a) < \kappa$, $j \neq i$, the difference between the two queues can never become larger than κ, so that the positive queue must also hit 0 at a finite time after Σ_q. Therefore, letting t_j denote the time at which queue j hits 0, $i = 1, 2$, we have

$$q_i(t) = 0 \text{ and } \dot{z}_{i,i}(t) = \lambda - z_{i,i}(t), \text{ for all } t > t_j \geq \Sigma_q^a. \text{ Furthermore, } t_j < \infty.$$

It follows that $z_{i,i}(t) \to \lambda$ as $t \to \infty$, so that $x^a(t) \to x_0^*$ as stated. □

It follows from (6.16) and Lemma D.1 that, if at the end of cycle we have $-\Delta^a(\Sigma_2^a) \leq \kappa$, then $\Sigma_q^a < \infty$ and $x^a(t) \to x_0^*$ as $t \to \infty$. In addition, $\Delta^a(t)$ was just shown to reach 0 in finite time, and $z_{1,2}^a$ and $z_{2,1}^a$ each reach 0 in finite time by construction. Therefore, $x_3^a(t)$ reaches $(0,0,0)$ in finite time. Using similar arguments to those in Theorem 5.2, we can prove that

LEMMA D.2. *x_0^* in (5.1) is the unique stationary point of the approximating system. Furthermore, if x_3^a does not oscillate indefinitely, then $x_3^a(t) = (0,0,0)$ for all large enough t, so that $x^a(t) \to x_0^*$ as $t \to \infty$.*

Lemmas D.1 and D.2 together complete the proof of Theorem 6.1.

D.2.2. *Proof of Theorem 6.2.* To study possible oscillatory behavior of the approximating system in (6.17) we use an iterative algorithm, similar to the one in §C.3.1, based on the arguments in §6.3.

An iterative algorithm for the approximating system. In the iterative algorithm each (half) cycle of x^a corresponds to an iteration. We use a superscript (k) denote the k^{th} iteration of the algorithm, and drop the superscript "a" for ease of notation, e.g., $T_1^{(1)}$ is the value of T_1^a in (6.9) in the first cycle of x^a, or equivalently, the first iteration of the algorithm.

We start by choosing a value $\Delta^{(0)} \equiv \Delta(0) > \kappa$ and use it to numerically compute $T_1^{(1)}$ via (6.9). The obtained value of T_1^a is then used to compute $\Delta^{(1)} \equiv \Delta^a(\Sigma_4^a) = -\Delta^a(\Sigma_2^a)$ via (6.15). We continue iterating this way until one of two things occur: either we see $\Delta^{(k)} > \kappa$ for all k or else we observe $\Delta^{(k)} \le \kappa$ for some $k \ge 1$, *in which case the algorithm is stopped.*

Similar to Lemma C.6 and Corollary C.5 we can show that there exists $\epsilon_\kappa^a > 0$ such that, if the algorithm can be iterated indefinitely, then $\Delta^{(k)} > \kappa + \epsilon_\kappa^a$ for all $k \ge 1$. Of course, for the approximating system we can characterize ϵ_κ^a explicitly, and its value can serve as an approximation for the value of ϵ_κ in Corollary C.5.

LEMMA D.3. *A necessary condition for endless oscillation is that, for all $k \ge 1$, $\Delta^{(k)} > \kappa + \epsilon_\kappa^a$, where $\epsilon_\kappa^a \equiv -\log(1-\tau)$. In particular, if $\kappa < \Delta^{(k)} < \kappa - \log(1-\tau)$ for some $k \ge 1$, then $\Delta^{(k+1)} < 0$, so that the algorithm is stopped.*

PROOF. For ϵ_κ^a in the statement of the lemma, assume that $\kappa < \Delta^{(k)} \le \kappa + \epsilon_\kappa^a$, for some $k \ge 1$. Then by (6.9)

$$T_1^{(k+1)} \le \frac{\kappa + \epsilon_\kappa^a - 1 + \mu - \kappa}{1+\mu} + \frac{1-\mu}{1+\mu} e^{-T_1^{(k+1)}} < \frac{\epsilon_\kappa^a - 1 + \mu}{1+\mu} + \frac{1-\mu}{1+\mu} < \frac{\epsilon_\kappa^a}{1+\mu}.$$

Therefore, $T_1^{(k+1)} < \epsilon_\kappa^a \equiv -\log(1-\tau)$. It follows from (6.16) that $\Delta^{(k+1)} < 0$. □

As was mentioned above, the approximating fluid model is a switching dynamical system with jumps. In this new setting, the approximating fluid solutions are elements in the space $\mathcal{D} \equiv \mathcal{D}[0,\infty)$ of real-valued right-continuous functions with limits everywhere, which we endow with the Skorohod J_1 topology, which we denote by d_t. Specifically, we consider the topological space $(\mathcal{D}, J_1)$, as in §3.3 of [48]. We have $x_k \to x$ in (D, J_1) as $k \to \infty$ if, for each t that is a continuity point of x,

$$d_t(x_k, x) \equiv ||x_k(\lambda_k(\cdot)) - x||_t \vee ||\lambda_k - e||_t \to 0 \quad \text{as} \quad n \to \infty,$$

where $e : [0,t] \to [0,t]$ is the identity function $e(s) \equiv s$, $0 \le s \le t$, λ_k is a homeomorphism of $[0,t]$ and $||\cdot||_t$ is the uniform norm applied to functions on the finite interval $[0,t]$. Note that convergence in J_1 reduces to uniform convergence over bounded intervals whenever the limit function is continuous, as is the case for all the solutions of (3.10).

We generalize Definition 4.4 by replacing the uniform metric in (4.1) with the Skorohod metric. We then say that a solution x^a spirals towards u^a_* if (4.1) holds for x^a and u^a_*, *but with the Skorohod J_1 metric replacing the uniform metric*. In our application we will let $\lambda_k(\Sigma_0^{(k)}) = \Sigma_0^{*(k)}$. After making that small perturbation of the switching times, so that they are aligned, we have uniform convergence over $[0,t]$.

The next lemma shows that spiraling of a solution x^a to u^a_* follows from the first limit in (4.1) and convergence of x^a to u^a_* at the four switching times. Its elementary proof is omitted.

LEMMA D.4. *Suppose that a periodic equilibrium u^a_*, having period T, exists for (6.17). If*

$$(I) \lim_{k\to\infty} T_i^{(k)} = T_i^* \quad \textit{and} \quad (II) \lim_{k\to\infty} \|x(\Sigma_i^{(k)}) - u(\Sigma_i^{*(k)})\| = 0, \quad 1 \le i \le 4,$$

for some solution $x^a \neq u^a_*$, then x^a spirals towards u^a_*. In particular,

$$\lim_{k\to\infty} d_t(x(\Sigma_0^{(k)} + \cdot), u^a_*(\Sigma_*^{(k)} + \cdot)) = 0,$$

for each continuity point t of $x(\Sigma_0^{(k)} + \cdot)$.

We are now prepared to prove Theorem 6.2 (a) and (b).

PROOF OF THEOREM 6.2 (*a*) AND (*b*). Lemma D.3 implies that a solution to the approximating system that oscillated indefinitely is bounded away from κ. Together with Lemma 6.2, this implies that $\Delta^{(k)}$ is confined

to the compact interval $I_\Delta \equiv [\kappa + \epsilon_\kappa^a, (1-\mu)(1-\tau)/\mu]$. Moreover, $\Delta^{(k)}$ is strictly monotone in $T_1^{(k)}$ by (6.16), which is itself strictly monotone in $\Delta^{(k-1)}$ by Lemma 6.1, $k \geq 1$. Hence, the sequence $\{\Delta^{(k)} : k \geq 0\}$ is monotone and bounded, and therefore converges to a limit $\Delta^{a,(\infty)} \in I_\Delta$. Since x_0^* is the unique stationary point of the approximating system and $\Delta^{a,(\infty)} > \kappa$ cannot be part of a stationary solution, the limit $\Delta^{a,(\infty)}$ must be a point on a periodic equilibrium, which is clearly unique. This proved (a). Part (b) of the theorem follows from Lemma D.4, together with Lemma D.3 and Theorem 6.1. □

D.2.3. *Proof of Theorem 6.2 (c).* It remains to show that the conditions of part (b) of Theorem 6.2 can be satisfied, i.e., there exist parameters for which $\Delta^{(k)} > \kappa$ for all $k \geq 0$ and $\Delta^{(k)} \to \Delta^{(\infty)} > \kappa$. To prove this, consider $\Delta^{(k-1)} > 1-\mu+\kappa$ and observe that, since $(1-\mu)/(1+\mu) < 1$, (6.9) implies that

$$0 < \frac{\Delta^{(k-1)} - 1 + \mu - \kappa}{1+\mu} < T_1^{(k)} < \frac{\Delta^{(k-1)} - 1 + \mu - \kappa}{1+\mu} + 1, \quad k \geq 1. \tag{D.1}$$

By Lemma 6.2, $\Delta^{(k-1)}$ is bounded from above by $\Delta_{bd}^a \equiv (1-\mu)(1-\tau)/\mu$. Therefore, consider $\Delta^{(0)} \in [\Delta_\mu^m, \Delta_\mu^M]$, where

$$\Delta_\mu^m \equiv 1 - \mu + \kappa \quad \text{and} \quad \Delta_\mu^M \equiv \Delta_{bd}^a \equiv (1-\mu)(1-\tau)/\mu. \tag{D.2}$$

Note that $\Delta_\mu^m > \kappa + \epsilon_\kappa^a$ for ϵ_κ^a in Lemma D.3 if τ is small, as we assume, and $1-\mu > \epsilon_\kappa^a$, which we require. The requirement that $\Delta_\mu^m < \Delta_\mu^M$, gives rise to quadratic equation in μ whose roots are

$$\begin{aligned} \mu_1 &= \frac{2 + \kappa - \tau - \sqrt{(\kappa-\tau)^2 + 4\kappa}}{2} \\ \text{and} \quad \mu_2 &= \frac{2 + \kappa - \tau + \sqrt{(\kappa-\tau)^2 + 4\kappa}}{2}, \end{aligned} \tag{D.3}$$

which are easily seen to satisfy $0 < \mu_1 < 1 < \mu_2$. Therefore, we henceforth consider $\mu \in (0, \mu_1)$ such that $1 - \mu > \epsilon_\kappa^a \equiv -\log(1-\tau)$, so that $\mu < 1 + \log(1-\tau)$.

Next, we introduce a mapping taking $\Delta(0) = \Delta$ into a function of T_1^a, where $T_1^a \equiv T_1^a(\Delta)$ is the unique positive solution to (6.9); specifically, let

$$\mathcal{T} : \Delta \mapsto -\kappa - \frac{1-\mu}{\mu} e^{-T_1^a} + \frac{1-\mu}{\mu}(1-\tau), \tag{D.4}$$

so that $\mathcal{T}(\Delta^{(k-1)}) = \Delta^{(k)}$, $k \geq 1$.

For fixed $\mu \in (0, \mu_1)$ and $0 < \delta_\mu < \Delta_\mu^M - \Delta_\mu^m$ to be specified below, let

$$\mathbb{S}_\mu \equiv [\Delta_\mu^M - \delta_\mu, \Delta_\mu^M]. \tag{D.5}$$

Note that the end points of $\mathbb{S}_\mu$ depend on μ, and that $\bigcup_\mu \mathbb{S}_\mu = [1+\kappa, \infty)$, where the union is taken over all the values of $\mu \in (0, \mu_1)$, for μ_1 in (D.3). In particular, the left end point of $\mathbb{S}_\mu$ is bounded from below whereas its right end point is unbounded as $\mu \downarrow 0$. Nevertheless, $\mathbb{S}_\mu$ is compact for any fixed $\mu \in (0, \mu_1)$.

LEMMA D.5 (sufficient condition for endless iterations). *For a given pair of control parameters (κ, τ) and μ_1 in (D.3), there exists $\mu_* \in (0, \mu_1)$ such that $\mathcal{T} : \mathbb{S}_\mu \to \mathbb{S}_\mu$ for all $\mu \le \mu_*$.*

PROOF. Observe that by (D.1) and (D.4)

$$\begin{aligned} \mathcal{T}(\Delta) &= -\kappa - \frac{1-\mu}{\mu} e^{-T_1^a} + \frac{(1-\mu)(1-\tau)}{\mu} \\ &> -\kappa + \frac{1-\mu}{\mu}(1 - \tau - e^{-\frac{\Delta - 1 + \mu - \kappa}{1+\mu}}), \end{aligned} \tag{D.6}$$

so that $\mathcal{T}(\Delta) > \Delta_\mu^m$ whenever the following inequality holds

$$\xi(\Delta) \equiv e^{-\frac{\Delta - 1 + \mu - \kappa}{1+\mu}} < 1 - \tau + \mu(1 - 2\kappa/(1-\mu)). \tag{D.7}$$

To see that (D.7) does hold for all $\mu \le \mu^*$, for some μ^* as in the statement of the lemma, observe that $\xi(\Delta)$ decreases to 0 as Δ increases to ∞ and that the right-hand side of (D.7) is bounded from below by $1 - \tau$ as μ decreases to 0. Since $\Delta_\mu^m \to 1 + \kappa$ and $\Delta_\mu^M \to \infty$ as $\mu \downarrow 0$, we can find μ_* small enough and Δ large enough such that, for all $\mu \le \mu_*$ and $\Delta_\mu^m < \Delta < \Delta_\mu^M$, (D.7) holds for that Δ.

Choose $c > 0$ such that $1 - \tau - c > 0$ and fix $0 < \epsilon < c$. Take μ_* smaller if needed, so that for any $\mu \in (0, \mu_*)$, it holds that $\xi(\Delta) < \epsilon$ whenever $\Delta > \frac{1-\mu}{\mu}(1 - \tau - c) - \kappa$. Then by (D.6) $\mathcal{T}(\Delta) > \frac{1-\mu}{\mu}(1 - \tau - \epsilon) - \kappa > \frac{1-\mu}{\mu}(1 - \tau - c) - \kappa$. The statement of the lemma follows by taking

$$\delta_\mu \equiv (1 - \mu_*)c/\mu_* + \kappa, \tag{D.8}$$

where we take μ_* sufficiently small to have $\Delta_\mu^M - \delta_\mu > \Delta_\mu^m$, i.e., $\frac{1-\mu}{\mu}(1 - \tau - c) - \kappa > 1 - \mu + \kappa := \Delta_\mu^m$. That is, $\mathcal{T}(\Delta) \in \mathbb{S}_\mu$ as stated. □

We use Lemma D.5 to obtain geometric rates of convergence to equilibrium in §H.

APPENDIX E: PROOF OF THE RESULTS IN SECTION 8

PROOF OF THEOREM 8.1. For each $n \geq 1$ consider the CTMC X^n initialized with its stationary distribution, namely, $X^n(0) \stackrel{\mathrm{d}}{=} X^n(\infty)$, $n \geq 1$. The sequence $X^n(\infty)$ is tight in $\mathbb{R}_6$ because each sequence of elements in the vector $\bar{X}^n$ is tight in $\mathbb{R}$. This follows immediately for $\bar{Z}^n_{i,j}(0)$, which are bounded from below by 0 and from above by some $c > 1$, $i, j = 1, 2$. Tightness of $\bar{Q}^n_1(0)$ and $\bar{Q}^n_2(0)$ follows from the infinite-server stochastic-order bound on the queues in Lemma A.5 in [31]. In particular, $\bar{Q}^n_i \leq_{st} \bar{Q}^n_{i,bd}$ pathwise, where $Q^n_{i,bd}$ is the number-in-system process in an $M/M/\infty$ queue with arrival rate λ^n_i and service rate θ. See also the proof of Theorem 8.2 where a similar bound is constructed.

By Theorem I.1, the sequence of processes $\{\bar{X}^n : n \geq 1\}$ is tight in $\mathcal{D}_6$, and we can therefore consider a converging subsequence of processes, whose initial conditions $\bar{X}^{n'}(0) \stackrel{\mathrm{d}}{=} \bar{X}^{n'}(\infty)$ also converge to some limit

$$\bar{X}(0) \equiv (\bar{Q}_i(0), \bar{Z}_{i,j}(0); i, j = 1, 2) \quad \text{in} \quad \mathbb{R}_6.$$

Since the initial condition is distributed according to the stationary distribution of $\bar{X}^n$, each of the CTMC's in the prelimit is stationary, and it follows that any limit of $\bar{X}^n$ must also be stationary process. In particular,

$$\bar{Z}_{i,j}(t) \stackrel{\mathrm{d}}{=} \bar{Z}_{i,j}(0) \quad \text{for all } t \geq 0 \quad \text{and} \quad (i,j) = (1,2) \quad \text{or} \quad (i,j) = (2,1).$$

First observe that, if $\bar{Z}_{1,2}(0) = \bar{Z}_{2,1}(0) = 0$ and $\bar{Q}_i(0) < \kappa$ w.p.1, then the two pools and their associated queues operate as two independent underloaded $M/M/m_i$ systems and therefore $\bar{X}(0) = x^*_0$ w.p.1, implying that $\bar{X}^n(\infty) \Rightarrow x^*_0$.

It follows from the routing rules of FQR-ART that for any sample path for which both $\bar{Z}_{1,2}(0)$ and $\bar{Z}_{2,1}(0)$ are strictly positive, at least one of these processes must be strictly decreasing over some interval $(0, \epsilon)$, $\epsilon > 0$, contradicting the stationarity of $\bar{X}$. Therefore, if $\bar{Z}_{i,j}(0) > 0$, then $\bar{Z}_{j,i}(0) = 0$, $i \neq j$ w.p.1.

Assume, for example, that $P(\bar{Z}_{1,2}(0) > 0) > 0$. Then there exists a measurable set $B_{1,2}$ in the underlying probability space, such that all the sample paths in $B_{1,2}$ have $\bar{Z}_{1,2}(0) > 0$ and $\bar{Z}_{2,1}(0) = 0$. Now, if $d_{1,2}(0) \neq 0$, where

$$d_{1,2}(t) \equiv \bar{Q}_1(t) - r\bar{Q}_2(t) - \kappa,$$

then $\bar{Z}_{1,2}$ is strictly increasing or strictly decreasing over some right neighborhood of 0, because $d_{1,2}$ is necessarily continuous by Theorem I.1. Hence, $d_{1,2}(t) = 0$, so that $q_1(t) \geq \kappa$ w.p.1 for all $t \geq 0$. In turn, $\bar{Z}_{1,1}(t) = m_1$

w.p.1 for all $t \geq 0$. However, this is impossible, because $\lambda_1 < \mu_{1,1} m_1$, so that $\bar{Q}_1(t)$ must be strictly decreasing if $\bar{Q}_1(0) > 0$. It follows that $P(B_{1,2}) = 0$. Symmetric arguments give that $P(\bar{Z}_{2,1}(0) > 0) = 0$ as well.

It follows that, if $\bar{Q}_i(0) > 0$, then $\bar{Q}_i$ must be strictly decreasing on some right neighborhood of 0, because $\bar{Z}_{i,i}(0) = m_i$. Hence, $\bar{Q}_i(0) = 0$. Then the X model is asymptotically two independent $M/M/n+M$ systems with service rate equals to 1 and arrival rate $\lambda^n < n$. Paralleling (C.1), we conclude that $\bar{X}(0) = x_0^*$ w.p.1, so that $\bar{X}^{n'}(\infty) \Rightarrow x_0^*$ as $n' \to \infty$. The statement of the theorem follows because the converging subsequence we considered was arbitrary. □

Proof of Theorem 8.2. Consider the queue process $Q_{bd}^n := \{Q_{bd}^n(t) : t \geq 0\}$ in an $M/M/\infty$ system that has arrival rate $2\lambda^n$ and service rate θ. Then Q_{bd}^n is distributed the same as the sum of the two queues in the X system in which the service process is "shut off" so that all the output from the two queues is due to abandonment. Specifically, we construct the X model and the $M/M/\infty$ system on the same probability space by giving both the same initial condition and the same Poisson arrival processes (exploiting the fact that a superposition of two independent Poisson processes is a Poisson process with the sum of the rates). If $Q_\Sigma^n(t) = Q_{bd}^n(t)$ and there is an abandonment from Q_Σ^n, then we can generate an abandonment from Q_{bd}^n; see, e.g., [47]. Therefore, Q_{bd}^n is never below Q_Σ^n.

It is well-known that the Markovian infinite-server queue is exponentially ergodic, see, e.g., Proposition 7.2 in [35]. However, we need to show that this implies that the same holds for X^n. We thus use the exponential drift condition on the generator of X^n whose state space is

$$\Xi \equiv \mathbb{Z}_+^2 \times \{0, 1, \ldots m^n\}^4.$$

For $x \in \Xi$, let $V(x) := (1+\gamma)^{x_1+x_2}$, for some $\gamma > 0$ which is characterized below. For Q_{bd}^n we consider the corresponding function $U(q) = (1+\gamma)^q$, $q = x_1 + x_2$. Then $V : \mathbb{R}_6 \to [1, \infty)$ is a *norm-like* function, namely $V(x) \to \infty$ as $\|x\| \to \infty$ (we use the standard norm on $\mathbb{R}_6$). Similarly, $U : \mathbb{R} \to [1, \infty)$ is a norm-like Lyapunov function for the generator of Q_{bd}^n.

Due to the sample-path stochastic order relation between Q_Σ^n and Q_{bd}^n, we have $\mathcal{Q}V \leq \mathcal{Q}_{bd}U$, where $\mathcal{Q}$ denotes the generator matrix of X^n and $\mathcal{Q}_{bd}$ denotes the generator matrix of Q_{bd}^n. Now, if we show that, for some compact set $C \subset \Xi$, the following exponential drift condition holds

$$\mathcal{Q}_{bd}U \leq -cV + d\mathbf{1}_C,$$

for strictly positive constants c and d and γ, then the statement of the theorem will follow from Theorem 2.5 in [22], because $\mathcal{Q}V \leq \mathcal{Q}_{bd}U$.

To that end, we recall that the off-diagonal components of $\mathcal{Q}_{bd}$ are given by

$$q_{i,i+1} = 2\lambda^n, \quad q_{i,i-1} = k\theta, \quad \text{and} \quad q_{i,j} = 0 \quad \text{for } |i-j| > 1, \qquad i \geq 1.$$

Then for $k \geq 1$

$$\begin{aligned}(\mathcal{Q}_{bd}U)(k) &= -\theta k\gamma[(1+\gamma)^{k-1} - (1+\gamma)^k] + 2\lambda^n[(1+\gamma)^{k+1} - (1+\gamma)^k] \\ &= -\gamma(1+\gamma)^{k-1}(\theta k - 2\lambda^n(1+\gamma)).\end{aligned}$$

The RHS in the above display is negative for all states k satisfying $\theta k - 2\lambda^n(1+\gamma) > 0$, or equivalently,

$$k > \frac{2\lambda^n}{\theta}(1+\gamma). \tag{E.1}$$

If $2\lambda^n/\theta \notin \mathbb{Z}_+$, then we can always choose $\gamma > 0$ small enough such that (E.1) holds for all $k \notin C \equiv \{0, 1, \ldots, \lceil 2\lambda^n/\theta \rceil\}$. Otherwise, if $2\lambda^n/\theta$ is an integer, we can simply make C larger, e.g., take $C \equiv \{0, 1, \ldots, 2\lambda^n/\theta + 1\}$, so that (E.1) holds for any state $k \notin C$ if $\gamma < \theta/2\lambda^n$. □

REMARK E.1. In general, the exponential drift condition in the above proof should hold for a "small set" C; see, e.g., [22]. In a discrete state space, as is the case here, any compact set is small.

APPENDIX F: CHECKING FOR CONGESTION COLLAPSE

When there is no abandonment, we cannot expect that the queues in an oscillating system will remain finite as time increases. Indeed, if

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t (z_{i,i}(s) + \mu z_{i,j}(s))ds < \lambda, \quad i,j = 1,2, \tag{F.1}$$

then the queues are not *rate stable*, i.e., the long-run average input rate λ is larger than the long-run average throughput rate, so that the queues will increase without bound. We now show how to estimate whether (F.1) holds.

In particular, we now show that the simplified heuristic approximation in §6.4 facilitates verification of (F.1) for a system that is known to converge to the unique periodic equilibrium. Let Σ_i^* and T_i^* denote the switching and holding times of the periodic equilibrium, $1 \leq i \leq 4$. Without loss of generality, consider pool 1. (Due to the symmetry, it is sufficient to check whether (F.1) holds for one of the pools.) Then, for

$$\zeta(s) \equiv z_{1,1}^a(s) + \mu z_{2,1}^a(s) = 1 - (1-\mu)z_{2,1}^a(s),$$

(F.1) becomes

$$L \equiv \lim_{t\to\infty} \frac{1}{t}\int_0^t \zeta(s)ds = \frac{1}{\Sigma_4^*}\int_0^{\Sigma_4^*} \zeta(s)ds$$
$$= \frac{1}{\Sigma_4^*}\left[\int_0^{T_1^*} \zeta(s)ds + \zeta(T_1^*)\int_0^{T_2^*} \zeta(s)ds + (\Sigma_4^* - \Sigma_2^*)\right],$$

where the first equality follows from the asymptotic periodicity of the solution, and the second equality follows from the symmetry of the model. Recall also that $z_{2,1}^a \equiv 0$, so that $z_{1,1}^a = 1$ over $[\Sigma_2^*, \Sigma_4^*]$, which gives the last term in the square brackets. We can use the last value of $\xi^{(k)}$ obtained from the algorithm above to serve as our approximation for $\xi^* \equiv \xi(\Delta_*^a)$, for $\xi(\cdot)$ in (D.7), together with (6.7) and (6.11) to approximate L.

Using the fact that $\Sigma_4^* = 2\Sigma_2^*$, we have (since $\Sigma_4^* - \Sigma_2^* = \Sigma_2^*$)

$$\begin{aligned} L &= 1 - \frac{1-\mu}{2\Sigma_2^*}\left[\int_0^{\Sigma_2^*} z_{2,1}^a(s)ds + \Sigma_2^*\right] \\ &\approx \frac{1+\mu}{2} - \frac{1-\mu}{2[-\log(\xi^*) + \log((1-\xi^*)/\tau)/\mu]} \\ &\times \left[\int_0^{-\log(\xi^*)} (1-e^{-s})ds + (1-\xi^*)\int_0^{\frac{\log\left(\frac{1-\xi^*}{\tau}\right)}{\mu}} e^{-\mu s}ds\right] \\ &= \frac{1+\mu}{2} - \frac{(1-\mu)[-\log(\xi^*) + \xi^* - 1 + (1+\xi^*-\tau)/\mu]}{2[-\log(\xi^*) + \log((1-\xi^*)/\tau)/\mu]}, \end{aligned} \tag{F.2}$$

with the approximation following by, first noting that $\Sigma_2^* = T_1^* + T_2^*$ and, second, replacing T_1^* and T_2^* with (6.18) and (6.19), respectively.

Note that, unlike the original system (3.10), in the approximating system we can first compute the periodic equilibrium, when it exists, via the iterative algorithm, and then check whether the system goes through congestion collapse. The heuristic approximation given here facilitates this inspection, via the computation in (F.2). More specifically, if a periodic equilibrium of (6.17) is found, and if this periodic equilibrium is associate with congestion collapse, then the queues necessarily increase to infinity as time increases, provided that x_3^a converges to u_*^a before either queue hits 0. We can then make sure that $\Sigma_q^a = \infty$ simply by initializing the two queues of the six-dimensional vector $x^a(0)$ at sufficiently large values, so that either queue does not reach state 0 during the first few cycles (i.e., before x_3^a is sufficiently close to u_*^a). Here, congestion collapse means that the queues will

have an increasing trend in the sense that each queue will be larger at the beginning of a cycle than its value at the beginning of the previous cycle. On the other hand, if the periodic equilibrium is not associated with congestion collapse, i.e., the total average service rate during the periodic cycle is smaller than the arrival rate, then the queues will have a decreasing trend, so that they must eventually reach 0, regardless of their initial condition. We conclude that there is no need to actually determine the exact values of the initial queue lengths, or to check wether $\Sigma_q^a = \infty$, but only to check wether a periodic equilibrium is associated with congestion collapse.

APPENDIX G: AN ALGORITHM TO COMPUTE THE HEURISTIC APPROXIMATION IN SECTION 6.4

We start by choosing a value $\Delta(0)$ such that $\xi^{(1)} \equiv \xi$ in (D.7) is sufficiently small (e.g., $\xi^{(1)} < 0.05$) and $T_1^{(1)}$ in (6.18) is strictly positive. Given $\xi^{(1)}$, we compute $\Delta^{(1)}(\Sigma_2^{(1)})$ in (6.20), and take $\Delta^{(1)}(0) = -\Delta^{(1)}(\Sigma_2^{(1)})$ in order to compute $\xi^{(2)}$ via (D.7). As before, we continue iterating until we see convergence to a legitimate value, i.e., $\Delta^{(k)}$ converges to some $\Delta_*^a > \kappa$ and $\xi^{(k)}$ converges to a value $\xi_* < 1$, or we obtain an illegitimate value at some iteration, i.e., $\Delta^{(k)} < \kappa$ or $\xi^{(k)} > 1$ for some $k \geq 1$. In the latter case, the algorithm is stopped. The latter case indicates that the solution x^a converges to x_0^*. If the initial condition for the algorithm is extreme, i.e., $\Delta^{(0)}$ is taken to be very large, then stopping the algorithm suggests that a periodic equilibrium does not exist.

APPENDIX H: STRONGER NOTIONS OF CONVERGENCE AND STABILITY

In Lemma D.5 we showed that for any κ and τ we can find μ_*, such that the iterative algorithm for the approximating system acts as a map from the space $\mathbb{S}_\mu$ in (D.5) into itself, thus ensuring that the algorithm can be iterated indefinitely. We now use Lemma D.5 and its proof to show that the iterative algorithm in §D.2.2 converges geometrically fast to the point Δ_*^a on the periodic equilibrium, when $u_*^a \in \mathbb{S}_\mu$. The fast monotone convergence to equilibrium is seen also in the numerical experiments in §7.

THEOREM H.1 (geometric rate of convergence). *Fix $c \in (0, 1-\tau)$ and consider $\mu \leq \mu_*$, for μ_* in Lemma D.5. Consider the solution x^a to the approximating system for a given initial condition $\Delta(0) = \Delta^{(0)} \in \mathbb{S}_\mu$. Then for any $\rho \in (0,1)$ there exists a $\mu_{**} \leq \mu_*$ such that, for all $\mu \leq \mu_{**}$ and δ_μ*

in (D.8),

$$|\Delta^{(k)} - \Delta^a_*| \le \frac{\rho^k}{1-\rho}|\Delta^{(1)} - \Delta^{(0)}| \le \delta_\mu \frac{\rho^k}{1-\rho}.$$

In particular, x^a_3 *converges to* u^a_* *geometrically fast in the number of cycles.*

Note that the statement of the theorem implies that there exists a unique asymptotically-stable periodic equilibrium in $\mathbb{S}_\mu$, as we already know.

PROOF. For any $\mu \le \mu_*$, $\mathcal{T}$ maps $\mathbb{S}_\mu$ into itself by Lemma D.5, in which case, for any $\Delta_1, \Delta_2 \in \mathbb{S}_\mu$, (D.4) gives

$$\begin{aligned} |\mathcal{T}(\Delta_1) - \mathcal{T}(\Delta_2)| &= \frac{1-\mu}{\mu} e^{\frac{1-\mu+\kappa}{1+\mu}} |e^{-\Delta_1/(1+\mu)} - e^{-\Delta_2/(1+\mu)}| \\ &\le \frac{1-\mu}{\mu} e^{\frac{1-\mu+\kappa}{1+\mu}} e^{-\frac{1-\mu}{\mu}(1-c)+\kappa} \frac{1}{1+\mu} |\Delta_1 - \Delta_2|. \end{aligned} \tag{H.1}$$

The inequality follows because, for $g(\Delta) \equiv e^{-\Delta/(1+\mu)}$,

$$|\dot{g}(\Delta)| \le K \equiv \frac{1}{1+\mu} e^{-\frac{1-\mu}{\mu}(1-c)+\kappa}, \quad \Delta \in \mathbb{S}_\mu \equiv [\Delta^M_\mu - \delta_\mu, \Delta^M_\mu],$$

for δ_μ in (D.8), implying that $g(\cdot)$ is Lipschitz continuous with a best Lipschitz constant that is no larger than K over the domain $\mathbb{S}_\mu$.

The RHS of the inequality in (H.1) clearly decreases to 0 as $\mu \downarrow 0$ for any two fixed Δ_1 and Δ_2. Hence, for any $\rho \in (0,1)$ we can find μ_{**} small enough, such that $|\mathcal{T}(\Delta_1) - \mathcal{T}(\Delta_2)| < \rho|\Delta_1 - \Delta_2|$ for all $\mu \le \mu_{**}$. In particular, if $\mu \le \mu_{**}$, then $\mathcal{T}$ is a contraction mapping from the compact interval $\mathbb{S}_\mu$ into itself.

Let $\mathcal{T}^{(k)}$ denote the k^{th} iteration of the map (D.4), i.e., $\mathcal{T}^{(k)} \equiv \mathcal{T} \circ \cdots \circ \mathcal{T}$, where the composition map $\circ$ is taken k times. Then $\mathcal{T}^{(k)}(\Delta^{(0)}) = \Delta^{(k)}$, $k \ge 1$, and the claim follows from the Banach fixed point theorem. □

By Lemma D.4, the three-dimensional solution x^a_3 to (6.17) "spirals" toward u^a_*. Using Theorem H.1, we next prove a stronger result, stating that the rate of convergence of an oscillating solution to the approximating system (in continuous time) is exponential.

Let $\mathcal{P}^a_*$ denote the image of the periodic equilibrium u^a_*;

$$\mathcal{P}^a_* \equiv \{\gamma \in \mathbb{S}^a : \gamma = u^a_*(t),\ 0 \le t < \Sigma^*_4\},$$

where $\mathbb{S}^a$ in §6 is the state space of the approximating system. Recall that the convergence of x^a_3 to u^a_* holds under the Skorohod metric defined in §D.2.2.

THEOREM H.2 (exponential stability). *Under the conditions of Theorem H.1* u_*^a *is exponentially stable, i.e., there exist constants* $\vartheta, \beta > 0$ *such that*

$$\inf_{u \in \mathcal{P}_*^a} \|x_3^a(\lambda(t)) - u\| < \vartheta e^{-\beta t}, \quad t \geq 0,$$

where $\lambda(\cdot)$ *is a homeomorphism of* $[0,t]$ *satisfying* $\lambda(\Sigma_0^{(k)}) = \Sigma_0^{*(k)}$ *for all* $k \geq 1$ *such that the* k^{th} *cycle falls in* $[0,t]$.

PROOF. It follows from Lemma D.4 and Theorem H.1 that, for all $k \geq 1$ and $t > \Sigma_*^{(k)}$,

$$\|x_3^a(\lambda(t)) - u_*^a(t)\| < \|x_3^a(\lambda(\Sigma_0^{(k)})) - u_*^a(0)\| \leq \frac{\|x_3^a(\Sigma_0^{(0)}) - u_*^a(0)\|}{1-\rho} e^{k \log(\rho)}.$$

Since x_3^a and u_*^a are uniformly bounded from above by Δ_μ^M in (D.2), the upper bound in (D.1) together with (6.13) give

$$\Sigma_2^{(k)} - \Sigma_0^{(k)} = T_1^{(k)} + T_2^{(k)} < \frac{\Delta_\mu^M - 1 + \mu - \kappa}{1+\mu} + 1 + \frac{\log(1/\tau)}{\mu} \equiv R,$$

so that $\Sigma_4^{(k)} - \Sigma_0^{(k)} < 2R$, for all $k \geq 1$. In particular, the length of any full cycle of any possible solution, including the periodic equilibrium, is smaller than $2R$. Since $\|x_3^a(\Sigma_0^{(0)}) - u_*^a(0)\| \leq \delta_\mu$, for δ_μ in (D.8), the statement of the theorem follows by taking $\vartheta \equiv \delta_\mu/(1-\rho)$ and $\beta \equiv -\log(\rho)/2R$. □

In ending we remark that the exponential bound on the rate of convergence to u_*^a should in general depend on the initial condition, as seen in the proof of Theorem H.2. In particular, exponential stability should in general be defined via $\|x_3^a(t) - u_*^a(t)\| < \vartheta \|x_3^a(0) - u_*^a(0)\| e^{-\beta t}$ for $\beta, \vartheta > 0$. However, we obtain the bound in the statement of the theorem since all the solutions we consider have values in $\mathbb{S}_\mu$, and are therefore uniformly bounded.

APPENDIX I: THE FLUID MODEL AS A LIMIT

The focus of the paper is on a fluid approximation for the stochastic X model under FQR-ART. In this section we prove that the switching fluid model arises as a many-server heavy-traffic fluid limit when a fluid-scaled sequence of these stochastic systems is considered. The proof of the *functional weak law of large numbers* (FWLLN) is given in §I.2, but we first expand on the stochastic model and many-server scaling in §I.1. We emphasize that, unlike the fluid limit proved in [31], the proof of the FWLLN here is standard because it does not include the stochastic averaging principle.

I.1. More on the stochastic model and heavy-traffic scaling. We now briefly expand on the review of the stochastic model, which was described in §2, and the heavy-traffic scalings. We consider a Markovian model, i.e., we assume that both arrival processes are independent (time-homogeneous) Poisson processes, and that service times, as well as patience times of customers waiting in queue, are exponentially distributed. Specifically, we assume that the class-i arrival rate in system n is λ_i^n, a class-i customer receives an exponentially-distributed service time in pool j with mean $1/\mu_{i,j}$, and a class-i customer has exponentially distributed patience with mean $1/\theta_i$, $i, j = 1, 2$. Customers who do not enter service before running out of patience will abandon the queue. (There is no abandonment from service.) All random variables are independent of each other and of the two arrival processes. Since FQR-ART is a Markovian control, in that the routing and scheduling decisions are a function of the state of the system and are independent of its history, it is easy to see that X^n in (2.2) is a six-dimensional time-homogeneous CTMC.

Due to abandonment of waiting customers, defining overloads is not entirely straightforward because a service pool can be considered normally loaded even if the traffic intensity to that pool is larger than 1. Our definition of overloads is taken from an asymptotic perspective. In particular, pool i is considered overloaded if $\rho_i > 1$, where

$$\rho_i \equiv \lim_{n\to\infty} \rho_i^n \equiv \lim_{n\to\infty} \lambda_i^n/(\mu_{i,i} m_i^n), \quad i = 1, 2.$$

On the other hand, we can have $\rho_i \leq 1$ with class i overloaded because there are many shared customers in pool i. This latter type of overload may be intentional, if sharing is deemed beneficial and is employed to alleviate an overload in the other class, or it may be caused by a harmful execution of the control, namely it is due to congestion collapse.

For any fixed n we must take $k_{1,2}^n$ to be sufficiently large so as to ensure that sharing begins only when the corresponding pool is genuinely overloaded due to a high arrival rate. In addition, $\tau_{1,2}^n$ should be sufficiently small to ensure that there is only a negligible amount of simultaneous two-way sharing. (Simultaneous sharing can occur because the direction of overload switches.) On the other hand, $\tau_{1,2}^n$ must be sufficiently large to be hit in a reasonable time. We refer to §§2.2 and 3.2 in [32] for elaborations on the reasonings behind the way we choose the thresholds. For our purposes here we simply enforce the following scaling assumption:

ASSUMPTION 3 (scaling parameters). *For strictly positive numbers m_i,*

λ_i, $k_{i,j}$ and $\tau_{i,j}$, $i,j = 1,2$,

$$m_i^n/n \to m_i, \quad \lambda_i^n/n \to \lambda_i, \quad k_{i,j}^n/n \to k_{i,j}$$
$$\text{and} \quad \tau_{i,j}^n/n \to \tau_{i,j} \quad \text{as} \quad n \to \infty.$$

Note that the first two limits in this assumption put us in the many-server heavy-traffic framework. The assumption that $\tau_{i,j} > 0$ will be relaxed for the approximating system for the fluid limit. See also J.1 below.

I.2. The FWLLN. Paralleling (3.1), we define for each $n \geq 1$

$$\mathcal{T}_1^n \equiv \inf\{t \geq 0 : Q_2^n(t) - rQ_1^n(t) \leq \kappa^n\}$$
$$\text{and} \quad \mathcal{T}_2^n \equiv \inf\{t \geq 0 : Z_{2,1}^n(\mathcal{T}_1^n + t) = \tau^n\}.$$

We also defined stopping times T_3^n, T_4^n and Σ_i^n, $1 \leq i \leq 4$ corresponding to the remaining holding times and switching times in (3.2).

Let

$$\Sigma_q^n := \inf\{t \geq 0 : \min\{Q_1^n(t), Q_2^n(t)\} = 0\}$$
$$\text{and} \quad \Sigma_q := \inf\{t \geq 0 : \min\{q_1(t), q_2(t)\} = 0\}.$$

As before, $\inf(\phi) \equiv \infty$. Since FQR-ART is non-idling, there cannot be any idleness in the system as long as both queues are strictly positive, i.e., if both queues are initially positive, then

$$Z_{1,1}^n(t) + Z_{2,1}^n(t) = Z_{2,2}^n(t) + Z_{1,2}^n(t) = n \quad \text{for all } t \leq \Sigma_q^n.$$

Notation. To present our results, we need to introduce some basic notation and refer to [48] for background. For $d \geq 1$, let $\mathcal{D}_d[0,t]$ denote the space of real-valued and right continuous $\mathbb{R}_d$-valued functions on an interval $[0,t] \subseteq \mathbb{R}_+$ that have limits from the left everywhere, endowed with the usual J_1 Skorohod topology. Let $\mathcal{C}_d[0,t] \subset \mathcal{D}_d[0,t]$ denote the (sub)space of $\mathbb{R}_d$-valued continuous functions defined on $[0,t]$. Recall that the J_1 topology is equivalent to the uniform topology in $\mathcal{C}_d(I)$ for any compact interval I. We use $\Rightarrow$ to denote convergence in distribution. We let e denote the identity function, $e(t) = t$, and $a \wedge b \equiv \min\{a,b\}$. Finally, we add a 'bar' to any fluid-scaled element (process or random variable), e.g., $\bar{X}^n \equiv X^n/n$.

THEOREM I.1 (FWLLN). *If $\bar{X}^n(0) \Rightarrow x(0)$ in $\mathbb{R}_6$ for some deterministic element $x(0) \in \mathbb{R}_6$ satisfying Assumption 2, then*

$$\bar{X}^n \Rightarrow x \quad \text{in} \quad \mathcal{D}_6[0, \Sigma_4 \wedge \Sigma_q \wedge t] \quad \text{as} \quad n \to \infty, \quad \text{for all } t \geq 0,$$

where x is a deterministic element of $\mathcal{C}_6$ and is the unique solution to the switching ODE $\dot{x} = f_\sigma(x)$, for f_σ in (3.10). Moreover,

$$n^{-1}(\mathcal{T}_i^n, \Sigma_i^n, \Sigma_q^n; 1 \le i \le 4) \Rightarrow (T_i, \Sigma_i, \Sigma_q; 1 \le i \le 4) \quad \textit{in} \quad \mathbb{R}_9 \quad \textit{as} \quad n \to \infty,$$

with $+\infty$ being a possible value as a limit of these stopping times.

By $+\infty$ being a possible value, e.g., $\Sigma_q^n \Rightarrow +\infty$, we mean that $P(\Sigma_q^n > M) \to 1$ as $n \to \infty$ for all $M > 0$.

Note that, if $x(0)$ satisfies Assumption 2, then necessarily $\Sigma_q > 0$. If, in addition, the fluid model is in the invariant set $\mathcal{O}$, then the convergence can be extended in an obvious way to any compact interval of $[0, \infty)$ because $\Sigma_q \equiv \infty$. Otherwise, $\Sigma_q < \infty$ and since $\lambda < 1$, class-i fluid will stop flowing to pool j, $i \neq j$. Since $P(|\Sigma_q^n - \Sigma_q| > \epsilon) \to 0$ as $n \to \infty$ (recall that convergence in distribution is equivalent to convergence in probability when the limit is deterministic), this show that sharing of customers will end at approximately time Σ_q in a large stochastic system.

The proof of Theorem I.1 follows standard pre-compactness arguments, combined with applications of the continuous-mapping theorem. We again refer to [48] for the general framework. We therefore start by representing the sample paths of X^n in terms of independent Poisson processes; see [27].

To simplify notation, let

$$\mathcal{A}_{1,2}^n(s) \equiv \{\{D_{1,2}^n(s) > 0\} \cap \{Z_{2,1}^n(s) \le \tau^n\}\}$$
$$\mathcal{A}_{2,1}^n(s) \equiv \{\{D_{2,1}^n(s) > 0\} \cap \{Z_{1,2}^n(s) \le \tau^n\}\},$$

LEMMA I.1 (martingale representation of X^n). *If $\min\{Q_1^n(0), Q_2^n(0)\} > 0$, then on the random interval $[0, \Sigma_q^n]$,*

$$\begin{aligned}
Q_1^n(t) &= M_1^n(t) + \lambda t - \int_0^t \theta Q_1^n(s) ds \\
&\quad - \int_0^t \mathbf{1}_{\mathcal{A}_{1,2}^n(s)} \left(Z_{1,1}^n(s) + \mu Z_{1,2}^n(s) + \mu Z_{2,1}^n(s) + Z_{2,2}^n(s)\right) ds \\
&\quad - \int_0^t (1 - \mathbf{1}_{\mathcal{A}_{1,2}^n(s)} - \mathbf{1}_{\mathcal{A}_{2,1}^n(s)}) \left(Z_{1,1}^n(s) + \mu Z_{2,1}^n(s)\right) ds, \\
Q_2^n(t) &= M_2^n(t) + \lambda t - \int_0^t \theta Q_2^n(s) ds \\
&\quad - \int_0^t \mathbf{1}_{\mathcal{A}_{2,1}^n(s)} \left(Z_{1,1}^n(s) + \mu Z_{1,2}^n(s) + \mu Z_{2,1}^n(s) + Z_{2,2}^n(s)\right) ds \\
&\quad - \int_0^t (1 - \mathbf{1}_{\mathcal{A}_{1,2}^n(s)} - \mathbf{1}_{\mathcal{A}_{2,1}^n(s)}) \left(Z_{2,2}^n(s) + \mu Z_{1,2}^n(s)\right) ds,
\end{aligned} \tag{I.1}$$

$$
\begin{aligned}
Z^n_{1,2}(t) &= M^n_{1,2}(t) + \int_0^t \mathbf{1}_{\mathcal{A}^n_{1,2}(s)} Z^n_{2,2}(s)ds - \int_0^t (1 - \mathbf{1}_{\mathcal{A}^n_{1,2}(s)})\mu Z^n_{1,2}(s)ds,\\
Z^n_{2,1}(t) &= M^n_{2,1}(t) + \int_0^t 1_{\mathcal{A}^n_{1,2}(s)} Z^n_{1,1}(s)ds - \int_0^t (1 - \mathbf{1}_{\mathcal{A}^n_{2,1}(s)}) Z^n_{2,1}(s))ds,\\
Z^n_{1,1}(t) &= n - Z^n_{2,1}(t),\\
Z^n_{2,2}(t) &= n - Z^n_{1,2}(t),
\end{aligned}
$$

where M^n_i and $M^n_{i,j}$, $i,j = 1,2$, are square-integrable martingales.

The expressions for all martingale terms in (I.1) can be inferred from (I.2) below. They are not presented explicitly since, as will be argued in the proof of Theorem I.1 below, they are asymptotically negligible under fluid scaling, and therefore play no role in the fluid limit.

PROOF. We use independent unit-rate Poisson processes to represent each of the component processes in (I.1). For example, the representation of Q^n_1 over $[0, \Sigma^n_q]$ is

$$
\begin{aligned}
Q^n_1(t) &= N^a_1\left(\lambda^n_1 t\right) - N^u_1\left(\theta_1 \int_0^t Q^n_1(s)ds\right)\\
&- N^+_1\left(\int_0^t \mathbf{1}_{\mathcal{A}^n_{1,2}(s)} \left(\mu_{1,1} Z^n_{1,1}(s) + \mu_{1,2} Z^n_{1,2}(s) + \mu_{2,1} Z^n_{2,1}(s) + \mu_{2,2} Z^n_{2,2}(s)\right) ds\right)\\
&- N^-_1\left(\int_0^t (1 - \mathbf{1}_{\mathcal{A}^n_{1,2}(s)} - \mathbf{1}_{\mathcal{A}^n_{2,1}(s)}) \left(\mu_{1,1} Z^n_{1,1}(s) + \mu_{2,1} Z^n_{2,1}(s)\right) ds\right),
\end{aligned}
$$

where N^a_1, N^u_1, N^+_1 and N^-_1 are mutually independent unit rate (homogeneous) Poisson processes.

Next, we exploit the fact that each of the Poisson processes in (I.1) minus its random intensity function constitutes a square-integrable martingale by Lemma 3.2 in [27], e.g.,

$$
M^{n,u}_1 \equiv N^u_1\left(\theta_1 \int_0^t Q^n_1(s)ds\right) - \theta_1 \int_0^t Q^n_1(s)ds \tag{I.2}
$$

is a square-integrable martingale. Thus, subtracting and then adding all the random intensities of the Poisson processes, and using the fact that a sum of martingales is again a martingale, we achieve the representation in the statement for Q^n_1 over the said interval. The representations for the other processes follow similar arguments. □

PROOF OF THEOREM I.1. Minor adjustments to the proof of Theorem 5.2 (and Corollary 5.1) in [31] give that $\{\bar{X}^n : n \geq 1\}$ is $\mathcal{C}$-tight in $\mathcal{D}_6$ with all limits being almost-everywhere differentiable. Those modifications to the aforementioned proof are straightforward, and are therefore omitted.

Next, by Doob's martingale inequality, the fluid-scaled martingales in (I.1) are asymptotically negligible, namely, $\bar{M}_i^n \Rightarrow 0e$ and $\bar{M}_{i,j}^n \Rightarrow 0e$ in $\mathcal{D}$, $i, j = 1, 2$, since these martingales are square integrable.

Given the initial condition, we have $\mathbf{1}_{\mathcal{A}_{1,2}^n(s)} = 0$ and $\mathbf{1}_{\mathcal{A}_{2,1}^n(s)} = 1$ over the interval $[0, \mathcal{T}_1^n \wedge \Sigma_q^n)$. Since any limit point of $\bar{X}^n$ is continuous, we must have that $P(\mathcal{T}_1^n \wedge \Sigma_q^n > \epsilon) \to 1$ for some $\epsilon > 0$. Therefore, it is easy to see from the representation of $\bar{X}^n$ with the indicator functions being constants over the interval $[0, \epsilon)$, that any limit point of $\bar{X}^n$ satisfies to the integral version of the ODE's in (3.4) and (3.5), whose unique solution implies that $\bar{X}^n$ converges to that solution x over $[0, \epsilon)$.

If $T_1 < \Sigma_q$, then the initial interval of convergence can be extended to $[0, T_1)$, and by Theorem 13.6.4 in [48], it holds that $\mathcal{T}_1^n \Rightarrow T_1$ in $\mathbb{R}$ as $n \to \infty$. Moreover, we have $\bar{X}^n(T_1) \Rightarrow x(T_1)$, so that $\mathbf{1}_{\mathcal{A}_{1,2}^n(s)} = \mathbf{1}_{\mathcal{A}_{2,1}^n(s)} = 0$ over the interval $[\mathcal{T}_1^n, (\mathcal{T}_1^n + \mathcal{T}_2^n) \wedge \Sigma_q^n)$ implies that

$$\lim_{n\to\infty} P(\mathbf{1}_{\mathcal{A}_{1,2}^n(s)} = \mathbf{1}_{\mathcal{A}_{2,1}^n(s)} = 0 \,;\; s \in (T_1, \Sigma_2 \wedge \Sigma_q) = 1.$$

Once again, plugging the constant values of the indicator functions to the representation (I.1) shows that any limit point of $\bar{X}^n$ satisfies the integral version of the ODE's in (3.7) and (3.8), whose unique solution on $[T_1, (T_1 + T_2) \wedge \Sigma_q)$ implies convergence of the sequence $\bar{X}^n$ to x. Moreover, we again have $\mathcal{T}_2^n \Rightarrow T_2$ in $\mathbb{R}$ as $n \to \infty$. Since T_1 and T_2 are deterministic, joint convergence of $(\mathcal{T}_1^n, \mathcal{T}_2^n)$ to (T_1, T_2) holds in $\mathbb{R}_2$ (e.g., Theorem 11.4.5 in [48]), so that $\mathcal{T}_1^n + \mathcal{T}_2^n \equiv \Sigma_2^n \Rightarrow \Sigma_2$ in $\mathbb{R}$ as $n \to \infty$.

The weak convergence of $\bar{X}^n$ to x and Σ_i^n to Σ_i can be extended to any compact subinterval of $[0, \Sigma_4 \wedge \Sigma_q]$ by exactly the same arguments. If $\Sigma_q > \Sigma_4$ we can then take $x(\Sigma_4)$ as a new initial condition and continue the proof inductively for all compact subinterval of $[0, \Sigma_q)$. □

APPENDIX J: IMPLICATIONS FOR THE CONTROL OF THE STOCHASTIC SYSTEM

J.1. Rescaling the thresholds.

Implications to the activation thresholds. As indicated in Assumption 3, the activation thresholds are asymptotically positive in fluid scale. This requires us to consider extreme cases with small abandonment rates and service rates for shared customers. In the *worst case* (leading to the biggest buildup of

queues) the abandonment rate is strictly smaller than the service rate of shared customers (and both are small).

For a given stochastic system there is freedom in choosing how to model the scaling of the thresholds. It is important that this freedom leads to ambiguities that must be accounted for. For example, if for $n = 100$, $m_1^n = m_2^n = 100$ and we take $k_{1,2}^n = k_{2,1}^n = 10$, then we can think of the activation thresholds as being equal to $0.1n$ or $\sqrt{n}$. From the fluid perspective, there are important difference between the two scalings. If the latter holds, then $\kappa = 0$ so that $\mathbb{S}_{1,2} = \mathbb{S}_{2,1}$ and the fluid model can cross from $\mathbb{S}_{1,2}^-$ to $\mathbb{S}_{2,1}^+$, and vice versa, in zero time. In this case, chattering and oscillations, as defined above, coincide, and are clearly more likely to occur. In particular, this suggests that oscillations can occur in the stochastic system even if a fluid approximation with $\kappa > 0$ does not oscillate at all, because a more appropriate approximation for the given system would be to assume that $\kappa = 0$; see Remark 5.1 below.

Implications to the release thresholds. There are important inconsistencies regarding the rescaling of the release thresholds. For example, in a system having 100 agents in each pool and arrival rate $\lambda^n = 98$, we may take $\tau_{i,j}^n = 3$. With these parameters, and regardless of the value of μ, pool j is clearly not overloaded at time t if $Z_{i,j}^n(t) \le \tau^n$, and the fluctuations of the queue must therefore be considered to be of order $o(n)$. However, the fluctuations of the queue will often be larger than τ^n, which is considered to be asymptotically positive under fluid scaling. Specifically, whereas

$$\|Q^n\|_T/\tau^n \Rightarrow 0 \text{ as } n \to \infty, \text{ for all } T > 0, \text{ where } \|Q^n\|_T \equiv \sup_{0 \le t \le T} Q^n(t),$$

we have $\|Q^n\|_T >> \tau^n$ for any reasonable value of n (which is not unrealistically large) and over intervals $[0, T]$, with $T = O(1)$ (e.g, $T \approx 1/\mu_{1,1}$.) It follows that, *relative to the stochastic fluctuations*, it is appropriate to think of the release thresholds as being $o(n)$ (even $O(1)$!). On the other hand, from a fluid-limit perspective, τ^n must satisfy Assumption 3, namely be strictly positive asymptotically in fluid scale, since otherwise $\bar{Z}_{i,j}^n := Z_{i,j}^n/n$ will not be hit this threshold in finite time when it is strictly decreasing; see §3.2 in [32].

We can think of the release thresholds as having a duality property in the fluid model: When $z_{i,j} \le \tau$ their affect on the system's performance is negligible, and we can consider them to be 0, i.e., $\tau_{i,j}^n = o(n)$. Whenever $z_{i,j} > \tau$ and is decreasing, we must think of τ as being strictly positive, so that τ^n is as in Assumption 3, to ensure that $z_{i,j}$ can hit τ in finite time. We take advantage of this duality property when constructing an approximation for the fluid model in §6.4.

Acknowledgements. We thank Editor Assaf Zeevi for suggestions that led to writing Section 9. The first author received support from National Science Foundation (NSF) Grant CMMI 1436518. The second author received support from NSF Grants CMMI 1265070 and 1634133.

REFERENCES

[1] Asmussen, S. (2003). *Applied probability and queues*. Springer. MR1978607
[2] Bell, S. L. and Williams, R.J. (2001). Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *The Annals of Applied Probability*, **11** (3) 608–649. MR1865018
[3] Billah, K. Y. and Scanlan, R. H. (1991). Resonance, Tacoma Narrows bridge failure, and undergraduate physics textbooks, *American Journal of Physics*, **59** (2), 118–124.
[4] Bramson, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, **30**, (1–2), 89–140. MR1663763
[5] Bramson, M. (1994). Instability of FIFO queueing networks. *The Annals of Applied Probability* **4** (2), 414–431. MR1272733
[6] Bramson, M. (2008). *Stability of queueing networks*. Springer. MR2445100
[7] Bramson, M. and Williams, R. J. (2000). On dynamic scheduling of stochastic networks in heavy traffic and some new results for the workload process *Proceedings of the 39th IEEE Conference on Decisions and Control*, 516–521.
[8] Chase, C., Serrano, J., Ramadge, P. J. (1993). Periodicity and chaos from switched flow systems: contrasting examples of discretely controlled continuous systems. *IEEE Transactions on Automatic Control* 38 (1), 70–83. MR1201496
[9] Dai, J. G. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *The Annals of Applied Probability* **5** (1) 49–77. MR1325041
[10] Durrett, R. (1991). *Probability: theory and examples*. Wadsworth and Brooks/Cole, Pacific Grove, California. MR1068527
[11] Erramilli, A. and Forys, L. J. (1991). Oscillations and chaos in a flow model of a switching system. *IEEE Journal on Selected Areas in Communications*, **9** (2), 171–178.
[12] Filippov, A. F. (1988) *Differential Equations with Discontinuous Righthand Sides*. Kluwer Academic Publishers, the Netherlands. MR1028776
[13] Garnett, O., Mandelbaum, A. and Reiman, M. (2002). Designing a call center with impatient customers, *Manufacturing & Service Operations Management*, **4**, (3), 208–227.
[14] Gurvich, I. and Whitt, W. (2009). Scheduling Flexible Servers with Convex Delay Costs in Many-Server Service Systems. *Manufacturing & Service Operations Management*, **11** (2), 237-253.
[15] Gurvich, I. and Whitt, W. (2009). Queue-and-Idleness-Ratio Controls in Many-Server Service Systems. *Mathematics of Operations Research* **34** (2), 363–396. MR2554064
[16] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29** (3) 567–588. MR0629195
[17] Harrison, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations *Stochastic differential systems, stochastic control theory and applications*, 147–186. MR0934722
[18] Harrison, J.M. (1998). Heavy traffic analysis of a system with parallel servers: asymp-

totic optimality of discrete-review policies. *Annals of applied probability* **8** (3) 822–848. MR1627791

[19] Harrison, J. M. and Taskar, M. I. (1983). Instantaneous control of Brownian motion. *Mathematics of Operations research*, **8** (3) 439–453. MR0716123

[20] Harrison, J. M. and Williams, R. J. (2005). Workload reduction of a generalized Brownian network. *The Annals of Applied Probability*, **15** (4), 2255–2295. MR2187295

[21] Khalil, H. K. (2002). *Nonlinear Systems.* Prentice Hall, New Jersey.

[22] Kontoyiannis I. and S.P. Meyn. (2003). Spectral Theory and Limit Theorems for Geometrically Ergodic Markov Processes. *The Annals of Applied Probability,* Vol. 13, 304–362. MR1952001

[23] Liberzon, D. (2003). *Switching in Systems and Control.* Birkhäuser, Boston. MR1987806

[24] Liu, Y. and Whitt, W. (2011). Nearly Periodic Behavior in the The Overloaded $G/D/S + GI$ Queue. *Stochastic Systems*, **1** (2), 340–410. MR2949544

[25] Lu S.H., Kumar P.R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Automat. Control.***36**(12), 1406–1416.

[26] Matveev, A. S., Savkin, A. V. (2000). *Qualitative theory of hybrid dynamical systems.* Birkhäuser, Boston.

[27] Pang, G., Talreja, R., Whitt, W., (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*, **4**, 193–267. MR2368951

[28] Perry, O., W. Whitt. (2009). Responding to unexpected overloads in large-scale service systems. *Management Sci.*, **55** (8), 1353–1367.

[29] Perry, O., W. Whitt. (2011a). A fluid approximation for service systems responding to unexpected overloads. *Oper. Res.*, **59** (5), 1159–1170. MR2864331

[30] Perry, O., W. Whitt. (2011b). An ODE for an overloaded X model involving a stochastic averaging principle. *Stochastic Systems*, **1** (1), 17–66. MR2948918

[31] Perry, O., W. Whitt. (2013). A fluid limit for an overloaded X model via an averaging principle. *Math, Oper. Res.*, **38** (2), 294–349. MR3062009

[32] Perry, O., W. Whitt. (2015a). Achieving rapid recovery in an overload control for large-scale service systems. Forthcoming in *INFORMS Journal on Computing*. Available at: `http://www.columbia.edu/~ww2040/allpapers.html` MR3382939

[33] Perry, O., W. Whitt. (2015b). Online Supplement to Achieving rapid recovery in an overload control for large-scale service systems. Forthcoming in *INFORMS Journal on Computing*. Available at: `http://www.columbia.edu/~ww2040/allpapers.html` MR3382939

[34] Reiman, M. I. (1984). Some diffusion approximations with state space collapse in *Modelling and performance evaluation methodology.* 207–240, Springer. MR0893658

[35] Robert, P. (2003). *Stochastic networks and queues*, Springer-Verlag. MR1996883

[36] Rudin, W. (1991). *Functional Analysis*, McGraw-Hill, Inc., New York. MR1157815

[37] Rybko A.N., Stolyar A.L. (1992). Ergodicity of stochastic processes describing the operations of open queueing networks. *Problems Inform. Transmission* **28**, 3–26 (in Russian). MR1189331

[38] Sastry, S.S. and Desoer, C. A. (1981). Jump behavior of circuits and systems. *IEEE Transactions on Circuits and Systems* **28** (12) 1109–1124. MR0643025

[39] Shah, D., D. Wischik. (2011). Fluid models of congestion collapse in overloaded switched networks. *Queueing Systems* **69** 121–143. MR2836736

[40] Schaft, V.D. and Schumacher, H. (2000). *An introduction to hybrid dynamical systems* Springer Lecture Notes in Control and Information Sciences, Vol. 251. Springer-Verlag, London. MR1734638

[41] Shakkottai, S, R. Srikant, A. L. Stolyar. (2004). Pathwise optimality of the exponen-

tial scheduling rule for wireless channels. *Adv. Appl. Prob.* **36** 1021–1045. MR2119854
[42] Stewart, D.E. (2000). Rigid-body dynamics with friction and impact. *SIAM review* **42** (1) 3–39. MR1738097
[43] Stolyar, A. L. (2004). Maxweight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. *Ann. Appl. Prob.* **14** (1) 1–53. MR2023015
[44] Teschl, G. (2009). *Ordinary Differential Equations and Dynamical Systems*, Universität Wien. Available online: http://www.mat.univie.ac.at/~gerald/ftp/book-ode/ode.pdf MR2961944
[45] Tezcan, T. and Dai, J. G. (2010). Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Operations Research*, **58** (1) 94–110. MR2668929
[46] Whitt, W. (1971). Weak Convergence Theorems for Priority Queues: Preemptive-Resume Discipline. *Journal of Applied Probability* **8** (1) 74–94. MR0307389
[47] Whitt, W. (1981). Comparing counting processes and queues. *Adv. Appl. Prob.* **13** (1) 207–220. MR0595895
[48] Whitt, W. (2002). *Stochastic-Process Limits*, New York, Springer. MR1876437
[49] Williams, R. J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing systems* **30** (1–2) 27–88. MR1663759

Ohad Perry
Department of Industrial Engineering and Management Sciences
Northwestern University
E-mail: ohad.perry@northwestern.edu

Ward Whitt
Department of Industrial Engineering and Operations Research
Columbia University
E-mail: ww2040@columbia.edu