INFORMS Journal on Computing
Vol. 27, No. 3, Summer 2015, pp. 491–506
ISSN 1091-9856 (print) | ISSN 1526-5528 (online)

http://dx.doi.org/10.1287/ijoc.2015.0642

Achieving Rapid Recovery in an Overload Control for Large-Scale Service Systems

Ohad Perry
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208,
ohad.perry@northwestern.edu

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027,
ww2040@columbia.edu

We consider an automatic overload control for two large service systems modeled as multiserver queues such as call centers. We assume that the two systems are designed to operate independently, but want to help each other respond to unexpected overloads. The proposed overload control automatically activates sharing (sending some customers from one system to the other) once a ratio of the queue lengths in the two systems crosses an activation threshold (with ratio and activation threshold parameters for each direction). In this paper, we are primarily concerned with ensuring that the system recovers rapidly after the overload is over, either because (i) the two systems return to normal loading or (ii) the direction of the overload suddenly shifts in the opposite direction. To achieve rapid recovery, we introduce lower thresholds for the queue ratios, below which one-way sharing is released. As a basis for studying the complex dynamics, we develop a new six-dimensional fluid approximation for a system with time-varying arrival rates, extending a previous fluid approximation involving a stochastic averaging principle. We conduct simulations to confirm that the new algorithm is effective for predicting the system performance and choosing effective control parameters. The simulation and the algorithm show that the system can experience an inefficient nearly periodic behavior, corresponding to an oscillating equilibrium (congestion collapse) if the sharing is strongly inefficient and the control parameters are set inappropriately.

Keywords: service systems; overload control; congestion collapse; time-varying queues; many-server queues; recover after overload incident; fluid models
History: Accepted by Winfried Grassmann, Area Editor for Computational Probability and Analysis; received August 2014; revised December 2014; accepted January 2015.

1. Introduction

An Automatic Overload Control. In this paper, we study an automatic control to temporarily activate "emergency" measures in an uncertain dynamic environment to mitigate damage from unexpected disruptions, and automatically return to normal operation when the how and when should the control be activated? And, second, how and when should the control be released?

As in our previous papers (Perry and Whitt 2009; 2011a, b; 2013; 2014) the specific setting we consider (described in detail in §2) involves two large-scale telephone call centers (or service pools within the same call center) that are designed to operate independently, but have the capability (due to network technology and agent training) to respond to calls from the other system, even though there might be some loss in service effectiveness and efficiency in doing so. These call centers are designed and managed to separately respond to uncertain fluctuating demand and, with good practices, can usually do so effectively; see Aksin et al. (2007) for background. However, these call centers may occasionally face exceptional unexpected overloads, due to sudden surges in arrivals, extensive agent absenteeism, or system malfunction (e.g., due to computer failures). It thus might be mutually beneficial for the two systems to agree to help each other during such overload incidents by serving some of the other customers.

In our previous papers, we proposed an automatic *fixed-queue-ratio-with-thresholds* (FQR-T) overload control for sharing when needed. While doing that work, we observed that our proposed FQR-T control needs to be further modified to ensure that the system recovers rapidly after an overload is over, either (i) because the two systems return to normal loading or (ii) because the direction of the overload suddenly shifts in the opposite direction; see Perry and Whitt (2011b, §2.2, paragraph 3) and Perry and Whitt (2013, Appendix B, Remark B.1). We contribute now by addressing that recovery problem. We do so by extending the previous FQR-T control to include

lower release thresholds (RT), producing a new *FQR control with activation-and-release thresholds* (FQR-ART). To study the system with the new FQR-ART control, we develop a new approximating six-dimensional deterministic fluid model, develop an efficient algorithm to analyze that fluid model, and perform simulation to verify the effectiveness of the control and the algorithm. We also generalize the previous models to incorporate the realistic feature of time-varying arrival rates and staffing functions. We are motivated by this call center application and queueing model, but the insights and analytical methods should be useful for other service systems and queueing models.

Congestion Collapse. An important feature of the FQR controls is that the sharing may be inefficient. A simple symmetric example that we will consider in §4 has identical service rates for agents serving their own customers, but identical slower service rates when serving the other customers. With such inefficiency, the whole system will necessarily operate inefficiently, with lower throughput of both classes, if both pools are busy serving the other customers instead of their own. Nevertheless, we find that judicious sharing with our proposed overload control can be effective even with some degree of inefficiency, but care is needed in setting the control parameters. A major concern with such inefficient sharing is that the system may possibly experience *congestion collapse*, i.e., the system may become overloaded due to the control, even though it has sufficient service capacity to handle all arrivals; see Erramilli and Forys (1991) and Shah and Wischiik (2011).

For the model considered here, we show in §4 that the two call centers can indeed experience behavior that is best described as congestion collapse if the sharing is strongly inefficient and an inappropriate control is used. An unstable oscillating equilibrium is predicted by our numerical algorithm for the approximating fluid model and confirmed by simulation; see Figures 3 and 4 later. We perform a detailed rigorous study of the challenging oscillatory behavior in Perry and Whitt (2015).

We emphasize that this oscillatory phenomenon is far from obvious because the stochastic model after the overload is over (without time-varying parameters) is an ergodic time-homogeneous continuous-time Markov chain (CTMC) with a steady-state limiting distribution. The situation that we consider in this paper is similar to the nearly periodic behavior of the $G/D/s + GI$ queue exposed in Liu and Whitt (2011). Here, by "nearly periodic," we mean that a periodic equilibrium exists for the fluid model, and that any oscillating fluid model will converge to that equilibrium in an appropriate sense as time increases. In particular, the fluid model does not have a unique steady state (fixed point). The reason for the discrepancy between the behavior of the stochastic system and its fluid model (which is the fluid limit of the system; see Perry and Whitt 2015) is that the two iterated limits (as time gets large and as the scale of the system gets large) done in a different order are not equal.

In this paper, we are primarily interested in identifying the possibility for congestion collapse due to oscillation, so that the control is designed appropriately to avoid this phenomenon. In particular, the fluid model and the algorithm for analyzing that model that we develop can be used to achieve the benefits of sharing while avoiding such bad behavior.

Congestion Collapse in the Queueing Literature. Within telecommunications there is a long history of congestion collapse and its prevention in the circuit-switched telephone network. More than 60 years ago, it was discovered that the capacity and performance of the network could greatly be expanded by allowing alternative routing paths; see Wilkinson (1956). If a circuit is not available on the most direct path, then the switch can search for free circuits on alternative paths. The difficulty is that these alternative paths may use more links and thus more circuits. This problem was first studied by simulation by Weber (1964). The classical remedy in such loss networks is trunk reservation control, where the last few circuits on a link are reserved for direct traffic; see Feinberg and Reiman (1994), Kelly (1991, §§4.3–4.5), and references therein.

Even though a call center can be regarded as a telecommunications network, our problem is quite different from the classical loss network setting discussed above, because queueing occurs and no customers are turned away. As a consequence, our system is more "sluggish;" it responds more slowly to changes in conditions, and presents new challenges. A thorough literature review is found in the online supplement (available as supplemental material at http://dx.doi.org/10.1287/ijoc.2015.0642) for this paper.

Organization of the Rest of the Paper. In §2, we define the stochastic X model and the FQR-T and FQR-ART controls. Building on simple fluid considerations, in §§3 and 4, we demonstrate the need to modify FQR-T to rapidly recover after the overload is over. In §3, we show why RTs are needed. In §4, we show that, unless precaution is taken, the RTs can cause congestion collapse when the system recovers from an overload. To avoid that bad behavior, the activation thresholds need to be increased beyond the FQR-T values. In §5, we develop the fluid approximation, and in §6, we develop an efficient algorithm to numerically solve it. Finally, in §7, we draw conclusions and suggest directions for further research.

Additional material appears in an online supplement. The supplement has an extended discussion about the related literature and our contribution to that literature; fluid models for an underloaded system (when at least one pool has idleness); three numerical examples that demonstrate the effectiveness of the FQR-ART control, and the fluid model by comparing the results of the numerical algorithm for the ordinary differential equation (ODE) to the results of simulation experiments; numerical and simulation examples that demonstrate congestion collapse due to oscillations; and finally, a fluid model for at least one pool is underloaded (has some fluid-scaled idleness).

2. The Time-Varying X Model

The X model has two customer classes and two agent pools, each with many homogeneous agents working in parallel. We assume that each customer class has a service pool primarily dedicated to it, but all agents are cross-trained so that they can handle calls from the other class, even though they may do so inefficiently, i.e., customers may be served at a slower rate when served in the other class pool. We assume that the service times are independent exponential random variables, with $1/\mu_{i,j}$ being the expected time for a class i customer to be served in service pool j. Each class has a buffer with unlimited capacity where customers who are not routed immediately into service upon arrival wait to be served. Within each class, customers enter service according to the first-come, first-served discipline. Customers have limited patience, so that they may abandon from the queue. The successive patience times of class i customers are independent and identically distributed (i.i.d.) exponential variables with mean $1/\theta_i$.

We assume that customers arrive according to independent *nonhomogeneous* Poisson processes, one for each class, with time-varying deterministic rate functions. The staffing levels are assumed to be time dependent as well, usually chosen to respond to anticipated changes in the arrival rates; see Liu and Whitt (2012b), and references therein. As discussed in §1, it is necessary to specify how the system responds when the staffing level of a service pool is scheduled to decrease. We too allow server switching (an agent can take over service from an agent scheduled to leave). Because service times are exponential, it thus suffices to let idle agents leave when staffing decreases, and the first agent to become idle leaves when all agents are busy when staffing is scheduled to decrease.

Even though we do not prove any limit theorems here, and instead develop direct fluid models to approximate the stochastic system, we will use asymptotic considerations in our analysis. We therefore consider a sequence of X systems, as just described, indexed by a superscript n. As is standard for many-server heavy traffic limits, the service rates and abandonment rates are independent of n, but the arrival rates and staffing levels increase. Specifically, for each $n \geq 1$, let $\lambda_i^n(t)$ be the arrival rate to pool i and let $m_j^n(t)$ be the number of agents in pool j at time t. For the fluid approximation, we assume that

$$\lambda_i^n(t)/n \to \lambda_i(t) \quad \text{and} \quad m_j^n(t)/n \to m_j(t) \quad \text{as } n \to \infty, \tag{1}$$

uniformly in t over each bounded time interval.

As in Liu and Whitt (2012a), we assume that the limit functions λ_i and m_j in (1) are piecewise smooth, by which we mean that they have only finitely many discontinuities in any finite interval, have limits from the left and right at each discontinuity point and are differentiable at all continuity points. That assumption is not restrictive for applications and supports analysis of the approximating fluid model by differential equations. For call center applications, it usually suffices to consider piecewise-constant functions, but we allow greater generality because our analytical methods can be applied in other settings, which will be shown in the following examples.

Let $Q_i^n(t)$ be the number of customers waiting in the class-i buffer and $Z_{i,j}^n(t)$ be the number of class-i customers in service pool j at time t in system n. Let the associated six-dimensional vector process be

$$X^n \equiv X^n(t) \equiv (Q_i^n(t), Z_{i,j}^n(t)\colon i, j = 1, 2), \quad t \geq 0. \tag{2}$$

To define asymptotic regimes, let $\rho_i^n(t) := \lambda_i^n(t)/(\mu_{i,i} m_i^n(t))$ be the instantaneous traffic intensity function of class i (and pool i) alone in system n at time t. By (1),

$$\rho_i^n(t) - 1 \to \beta_i(t) \quad \text{as } n \to \infty, \tag{3}$$

uniformly in t over each bounded time interval. We say that class i (and pool i) is *underloaded* at time t if $\beta_i(t) < 0$, *overloaded* at time t if $\beta_i(t) > 0$, and *normally loaded* at time t if $\beta_i(t) = 0$.

The generality we have introduced allows for many possible scenarios, but here we restrict attention to an unexpected overload incident followed by a subsequent instantaneous switch in state, either (i) a return to normal loading or (ii) a switch in the direction of overloading. Thus, there are now three intervals: first normally loaded, then overloaded, and then a final new regime, which is either normal loading for both classes or an overload in the opposite direction. During each of these three intervals, the arrival rates and staffing functions are allowed to change.

Like before, we consider the system starting at the unanticipated time when the first overload incident begins. However, now the arrival rates and staffing functions no longer need to be constant within each

interval. By assumption, they have discontinuities at the beginning of the first overload incident and at the subsequent time when the overload is over. For the generality that we do consider, we exploit the fact that we know how to staff to stabilize the fluid system in face of time-varying arrival rates under normal loading; see Liu and Whitt (2012a, b), and references therein.

We assume that overloads may occur at an unanticipated time due to a sudden shift of the arrival rates to new and unknown values, or due to an unplanned decrease in the total service rates (e.g., due to agent absenteeism). One can then model the problem as a hybrid stochastic system (X^n, p), where X^n is defined in (2) and p represents the environment, e.g., p may achieve the values $0, 1, 2$, where $p = 0$ represents normal loads, and $p = i$ stands for a class-i overload $i = 1, 2$. However, this approach requires tracking the arrival and service rates continuously, which may be hard to do in practice. Instead, we propose a *state-dependent control* that can effectively respond to changes in the system's loads, and under which X^n is a nonhomogeneous CTMC.

2.1. The Initial FQR-T Control

Before describing the control, we review the original FQR-T control and demonstrate why it must be adjusted for the time-varying setting considered here. For each $n \geq 1$, the FQR-T control is based on two positive (activation) thresholds, $k^n_{1,2}$ and $k^n_{2,1}$ and the two queue-ratio parameters, $r_{1,2}$ and $r_{2,1}$ (which are chosen independent of n under (1)). We define the following two (centered) queue-difference stochastic processes:

$$D^n_{1,2}(t) \equiv Q^n_1(t) - k^n_{1,2} - r_{1,2}Q^n_2(t) \quad \text{and}$$
$$D^n_{2,1}(t) \equiv r_{2,1}Q^n_2(t) - k^n_{2,1} - Q^n_1(t), \quad t \geq 0. \tag{4}$$

As long as $D^n_{1,2}(t) < 0$ and $D^n_{2,1}(t) < 0$, we consider the system to be *not overloaded* so that no customers are routed to be served in the other class pool. Indeed, if the activation thresholds are chosen appropriately, then the event $\{D^n_{i,j}(t) \geq 0\}$ will occur with probability converging to 0 as $n \to \infty$ *unless* $\beta_i > 0$; we elaborate below. Once one of these inequalities is violated, the system is considered to be overloaded, and sharing is initiated. For example, if $D^n_{1,2}(t) \geq 0$, then class 1 is judged to be overloaded (because then $Q^n_1 - r_{1,2}Q^n_2 \geq k^n_{1,2}$), and it is desirable to send class-1 customers to be served in pool 2. Note that $D^n_{1,2}(t) \geq 0$ does not exclude the case that class 2 is also overloaded; we can have $\beta_i(t) > 0$ for *both* i. However, once one of the thresholds is crossed, its corresponding class is considered to be "more overloaded" than the other class. We refer to this situation as *unbalanced overloads*. We call $k^n_{1,2}$ and $k^n_{2,1}$ *activation thresholds*, because exceeding one of these thresholds activates sharing (and not exceeding prevents sharing when it is not desired).

The behavior of X^n in (2) depends on the choice of the thresholds $k^n_{i,j}$. In particular, we want the thresholds to be large enough so that sharing will not take place if both service pools are normally loaded, and small enough to detect any overload quickly, and start sharing in the correct direction once the overload begins. Note that without sharing, the two pools operate like two independent $M_t/M/m^n_t + M$ (time-varying Erlang-A) models. The familiar fluid and diffusion limits for the stationary Erlang-A model give insight as to how to choose these thresholds; e.g., see Garnett et al. (2002) and Pang et al. (2007). In Perry and Whitt (2013, Assumption 2.4) and Perry and Whitt (2014, Assumption 3), we assumed that the activation thresholds are chosen to satisfy

$$k^n_{i,j}/n \to 0 \quad \text{and} \quad k^n_{i,j}/\sqrt{n} \to \infty \quad \text{as } n \to \infty,$$
$$i, j = 1, 2 \text{ with } i \neq j. \tag{5}$$

The first limit in (5) ensures that overloads are detected quickly (immediately in the fluid model obtained as $n \to \infty$), whereas the second limit in (5) ensures that stochastic fluctuations of normally loaded pools will not cause undesired sharing, since the diffusion-scaled queue in that case are of order $\sqrt{n}$.

Given that the system is designed so that sharing of customers takes place only during overloads, it is reasonable to assume that agents serve the other class customers (the so-called shared customers) at a slower rate than they serve their own designated customers. Thus, substantial sharing is likely to reduce the effective service rate of the helping pool. In our previous work, we took measures to avoid sharing in both directions simultaneously. In particular, we imposed the one-way sharing rule described in §1. However, it is evident that the one-way sharing rule may considerably slow the recovery after the overload is over. We elaborate in §3.

To remedy this problem, we could consider removing the one-way sharing rule altogether and rely solely on the activation thresholds to avoid undesired sharing. However, removing the one-way sharing rule makes it necessary to increase the activation thresholds substantially, increasing the time until overloads are detected. Moreover, if these thresholds are too large, then some overloads may not be detected at all, because abandonment keeps the queues from increasing indefinitely. (Whereas there is also a need to increase the activation thresholds in our setting here, that increase is less than would be required if the one-way sharing was completely removed.) Moreover, if sharing is taking place in one direction and then immediately starts in the other direction in response

to a switch in the overload, then the combined service capacity of both pools may be reduced significantly, creating a period of severe congestion in both directions. Hence, it is beneficial to avoid too much simultaneous two-way sharing; see Perry and Whitt (2009, Example 2). Therefore our new control relaxes the one-way sharing rule by introducing the RTs alluded to earlier. We elaborate in the following section.

2.2. The Proposed FQR-ART Control

For the reasons discussed, we suggest a modification of the one-way sharing rule by introducing RTs. For each $n \geq 1$, we introduce two strictly positive numbers $\tau^n_{1,2}$ and $\tau^n_{2,1}$. A newly available type-2 agent is allowed to take a class-1 customer at time t only if $Z^n_{2,1}(t) \leq \tau^n_{2,1}$, i.e., if the number of type-1 agents serving class-2 customers at the same time t is below $\tau^n_{2,1}$ (and, of course, $D^n_{1,2}(t) \geq 0$), and similarly in the other direction. Ways to choose the parameters $\tau^n_{1,2}$ and $\tau^n_{2,1}$ will be discussed later.

However, the new RTs allow small simultaneous sharing in both directions, which can slightly increase the overload. In some cases, this slight increase in the overload is sufficient to cause the system to spin out of control and start to oscillate. The study of that oscillatory behavior is challenging, and is therefore rigorously established separately in Perry and Whitt (2015). Here, we demonstrate these oscillations in a simulation example; see §4. In particular, the new RTs make activation thresholds satisfying (5) unsuitable. We therefore conclude that these activation thresholds should be positive in "fluid scale," i.e., they should be chosen to satisfy

$$\lim_{n\to\infty} k^n_{i,j}/n = k_{i,j} > 0, \quad i,j = 1,2. \tag{6}$$

Thus, the FQR-ART control is specified by the parameter six-tuple $(r_{1,2}, r_{2,1}, k^n_{1,2}, k^n_{2,1}, \tau^n_{1,2}, \tau^n_{2,1})$ and the routing and scheduling rules that depend on the values of the two processes, $D^n_{i,j}$ and $Z^n_{i,j}$, $i \neq j$, in the manner described above. Note that FQR-T requires knowing only the queue lengths $Q^n_i(t)$ at each time t (specifically, the values of the two difference processes (4)), whereas FQR-ART also requires knowledge of $Z^n_{1,2}$ and $Z^n_{2,1}$. Under either control, the X model is a (possible inhomogeneous) CTMC.

2.3. Analysis via Fluid Approximations

Since the stochastic process X^n in (2) under FQR-ART is evidently too difficult to analyze exactly, we will employ a deterministic dynamical system approximation, and refer to that approximation as "fluid approximation" or "fluid model" interchangeably. The main idea in using fluid approximations is that, for large n, $\bar{X}^n \approx x$, for some deterministic function x that is easier to analyze than the untractable stochastic process X^n. We use the "bar" notation throughout to denote fluid-scaled processes, e.g., $\bar{X}^n \equiv X^n/n$. In particular, the fluid counterpart of X^n in (2) is the six-dimensional deterministic function

$$x \equiv x(t) \equiv \big(q_i(t), z_{i,j}(t)\colon i,j = 1,2\big), \quad t \geq 0,$$

where q_i and $z_{i,j}$ are the fluid approximations for the stochastic processes Q^n_i and $Z^n_{i,j}$, $i,j = 1,2$. The approximation $\bar{X}^n \approx x$ should be supported by a *functional weak law of large numbers* (FWLLN), stating that $\bar{X}^n \Rightarrow x$ as $n \to \infty$, extending Perry and Whitt (2013), but that remains to be established. (However, the FWLLN has been established for the FQR-T model in Perry and Whitt 2013.)

The value of the state-dependent control is also apparent in the fluid approximation, since, if we were to consider a stochastic hybrid system (X^n, p), $p \in \mathcal{P} := \{0,1,2\}$, as described above, then the fluid approximation would be a hybrid dynamical system of the form $\dot{x} = \Psi(x,p)$, for some function $\Psi\colon \mathbb{R}_6 \times \mathcal{P} \to \mathbb{R}_6$ that is discontinuous in its first argument due to (i) the switching of the dynamics caused whenever the value of p changes and (ii) state space collapse (SSC) when sharing takes place to keep the two queues at their designated ratio. Nevertheless, developing the fluid model for the system under FQR-ART is still challenging because of the need to "translate" the control in the stochastic system to a control in the deterministic dynamical system.

In the stochastic system, customer routing depends on the values of the difference processes in (4). For example, if sharing is taking place with pool 2 helping class 1, and assuming $Z^n_{2,1} \leq \tau^n_{2,1}$, the process $D^n_{1,2}$ determines which customer class a newly available type-2 agent will take. As shown in Perry and Whitt (2011a, b; 2013), that convention implies that the resulting fluid model is much more complicated than most fluid models in the literature. In particular, in the fluid system we cannot simply replace the process $D^n_{1,2}$ with its fluid counterpart process $d_{1,2}(t) \equiv q_1(t) - k_{1,2} - r_{1,2}q_2(t)$, $t \geq 0$ when sharing takes place. In fact, the purpose of the control is to produce SSC by keeping $d_{1,2}$ fixed at 0 during the overload. (Note that, in that case, the value of the three-dimensional process $(q_1, z_{1,2}, z_{2,1})$, say, determines the value of the six-dimensional process x, implying that SSC indeed occurs.) Hence, a refined asymptotic analysis of the behavior of $D^n_{1,2}$ (or $D^n_{2,1}$ during overloads in the other direction) is required. That refined analysis can be carried out thanks to a stochastic average principle (AP), which replaces the processes $D^n_{i,j}$, $i,j = 1,2$, with the long-run average behavior of corresponding limiting stochastic processes. In turn, those deterministic long-run averages determine the evolution of the fluid model; see §5, where the fluid equations are developed.

3. The Need to Relax the One-Way Sharing Rule

Relying on the fluid approximation, we now demonstrate why the one-way sharing rule impedes recovery after the overload incident is over. The simple fluid analysis suggests that RTs provide a good remedy, and helps indicate how they should be chosen.

3.1. The Recovery Time with One-Way Sharing

We consider two consecutive time intervals $I_1 = [t_0, t_1)$ and $I_2 = [t_1, t_2)$ with $0 \le t_0 < t_1 < t_2 \le \infty$, with the system being overloaded in opposite direction over each interval. Suppose that class 2 is overloaded over the time interval I_1 and that sharing is taking place with pool 1 helping class 2. Then, at time t_1, the loads suddenly change in such a way that sharing is required in the other direction. In particular, we assume that $\beta_1(t) \le 0$ and $\beta_2(t) > 0$ for $t \in I_1$, whereas $\beta_1(t) > 0$ and $\beta_2(t) \le 0$ for $t \in I_2$. We also assume that $z_{2,1}(t_1) > 0$.

We do two different mathematical analyses. We first consider a direct fluid model analysis, and then afterward, we consider the stochastic system. A fluid approximation for the evolution of $Z^n_{1,2}$ (which we refer to as $z_{1,2}(t)$) can easily be derived using rate considerations. Since every type-1 agent who is helping a class-2 customer at time $t > t_1$ will finish service immediately after time t at a rate $\mu_{2,1}$, regardless of the value of t, due to the memoryless property, and since there are no more class-2 customers routed to pool 1 after time t_1, we expect that $z_{2,1}$ will satisfy the ODE $\dot{z}_{2,1}(t) = -\mu_{2,1} z_{2,1}(t)$, $t \in I_2$, whose unique solution is

$$z_{2,1}(t) = z_{2,1}(t_1) e^{-\mu_{2,1} t}, \quad t \in [t_1, t_2). \tag{7}$$

As a consequence, for the fluid model, if $z_{2,1}(t_1) > 0$, then pool 1 will *never* empty, so that sharing can *never* begin in the opposite direction.

We now characterize the random time T^n after the time t_1 in the stochastic system with scale n for $Z^n_{2,1}(t)$ to first hit 0. The time required for all these customers to complete service is the maximum of $Z^n_{2,1}(t_1)$ i.i.d. exponential random variables. It is well known that the maximum of n i.i.d. exponential random variables with mean 1 is the harmonic sum $H_n \equiv \sum_{j=1}^n (1/j)$. Moreover, it is well known that $H_n - \log_e n \to \gamma$ as $n \to \infty$, where $\gamma \equiv 0.57721\ldots$ is the Euler-Mascheroni constant, e.g., see Young (1991). This limit is relevant for us, because from the established FWLLN in Perry and Whitt (2013), we know that having $z_{2,1}(t_1) > 0$ implies that $Z^n_{2,1}(t_1) \approx z_{2,1}(t_1) n$.

Hence, given $Z^n_{2,1}(t_1)$ and its approximate value, for large n,

$$E[T^n] = \sum_{j=1}^{Z^n_{2,1}(t_1)} \frac{1}{j \cdot \mu_{2,1}} \approx \frac{\log_e (Z^n_{2,1}(t_1))}{\mu_{2,1}} \approx \frac{\log_e (n z_{2,1}(t_1))}{\mu_{2,1}}. \tag{8}$$

We thus see that the expected time required for a pool to empty its shared customers after an overload is over, and no new shared customers are routed to that pool, is of order $\log_e(n)$ as $n \to \infty$.

3.2. Choosing Appropriate Release Thresholds

The simple considerations leading to (7) and (8) show that a large system will be slow to recover after an overload is over. That analysis also helps choose appropriate RTs. Indeed, the fluid model easily generates an approximate recovery time. In particular, if a RT of $\tau_{2,1}$ is used in the fluid model starting with $z_{2,1}(t_1)$ at time t_1, where $z_{2,1}(t_1) > \tau_{2,1} > 0$, then the RT will be hit at time

$$T \equiv \frac{1}{\mu_{2,1}} \log_e \left(\frac{z_{2,1}(t_1)}{\tau_{2,1}} \right).$$

This analysis indicates that the RTs in stochastic system n should be of order $O(n)$ as n increases. It suffices to pick two strictly positive numbers $\tau_{1,2}$ and $\tau_{2,1}$, and let

$$\tau^n_{1,2} \equiv n\tau_{1,2} \quad \text{and} \quad \tau^n_{2,1} \equiv n\tau_{2,1}. \tag{9}$$

With the scaling in (9), the recovery time T^n in system n should be approximately a constant, independent of n.

In summary, with FQR-ART, an available type-2 agent is allowed to serve a class-1 customer only if $Z^n_{2,1}(t) \le \tau^n_{2,1}$ (or, equivalently, only if $\bar{Z}^n_{2,1}(t) \le \tau_{2,1}$), and, of course, $D^n_{1,2}(t) \ge 0$, and similarly in the other direction. The choice in (9) shows that the RTs should be proportional to n, but does not determine the proportionality constants, $\tau_{1,2}$ and $\tau_{1,2}$. Further analysis shows that these can be quite small, as we show next.

3.3. Simulation Experiments

To illustrate the importance of the RTs for stochastic systems, we conducted simulation experiments, comparing the performance of a system with and without RTs. The results can be seen in Figures 2 and 3. The (fixed) parameters for this simulation are $m^n_1 = m^n_2 = 1{,}000$, $\lambda^n_1 = 1{,}200$, $\lambda^n_2 = 990$, $\mu_{1,1} = \mu_{2,2} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.5$, $\kappa^n_{1,2} = \kappa^n_{2,1} = 100$, $r_{1,2} = r_{2,1} = 1$.

Here, we can think of n as being fixed and equal to 1,000. With these parameters, $\rho^n_1 = 1.2$ and $\rho^n_2 = 0.99$, where $\rho^n_i \equiv \lambda^n_i / (m^n_i \mu_{i,i})$, so that class 1 may be regarded as overloaded, whereas class 2 may be regarded as normally loaded (recall (3)).

To respond to that unbalanced overload by having pool 2 help class 1, we should have $Z^n_{1,2} > 0$ and $Z^n_{2,1} = 0$ if one-way sharing is employed. However, we initialize the system at time 0 sharing in the opposite direction, with *all* pool-1 agents serving class-2 customers. We are interested in the time it takes the stochastic process $Z^n_{2,1}$ to reach 0, so that the desired

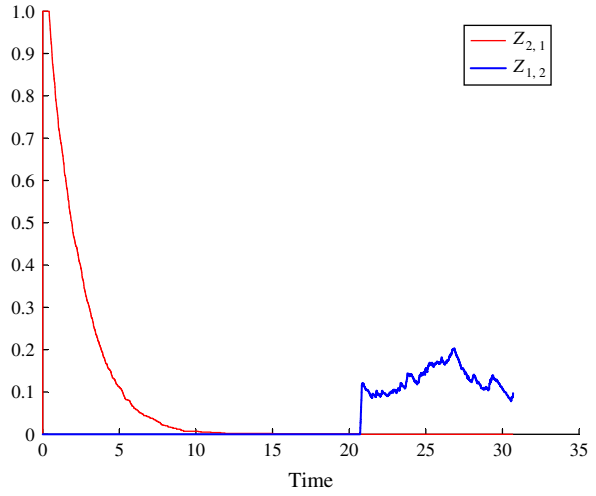

Figure 1 (Color online) Shared Customers When Initialized Incorrectly without any Release Thresholds

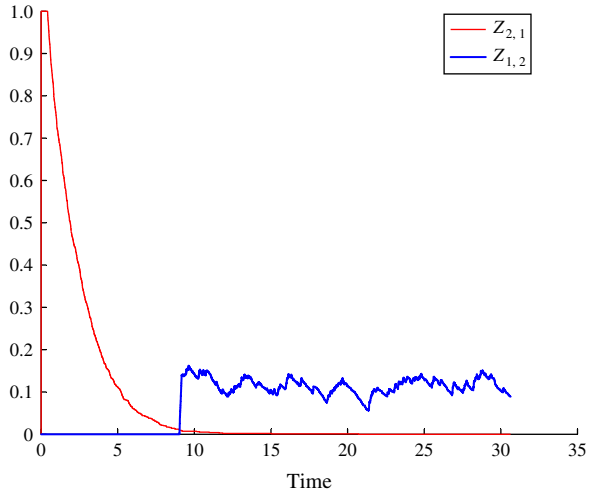

Figure 2 (Color online) Shared Customers When Initialized Incorrectly, with Release Thresholds $\tau_{1,2} = \tau_{2,1} = 0.01$

sharing can begin. Without RTs, the required recovery time is quite long, approximately 21 (mean service times of their own type). In contrast, with RTs of only $\tau^n_{1,2} = \tau^n_{2,1} = 0.01n = 10$, that time is reduced from about 21 to 9 service times. Thus, clearing the last 1% of the class-2 customers in pool 1 without RTs takes more than half the total clearing time.

We hasten to admit that we just considered an extreme example in which *all* service pool 1 is initially busy with customers from class 2. We did so to convey the message that *it is the last few agents working with class* 1 *that cause the largest part of the delayed response*. In particular, the $Z^n_{2,1}$ process decreases fast at the beginning, but then the decrease rate slows down considerably.

From Figures 1 and 2, it is also easy to see what happens in less extreme cases, when $0 < Z_{2,1}(0) < m_1$. For example, if we initialize with 20% sharing in the wrong direction, we see that, without a RT, the time to activate sharing in the right direction is about $21 - 4 =$ 17 time units. In contrast, with RTs, it is about $9 - 4 =$ 5 time units. Figures 1 and 2 show that the common value 4 in these calculations is the time to go from 100% sharing in the wrong direction to only 20% sharing in the wrong direction, which would be the same in the two cases. When we start with a lower percentage of agents sharing the wrong way, the difference becomes even more dramatic, because we eliminate a common initial period (here of length 4 time units).

4. Congestion Collapse Due to Oscillations

Section 3 dramatically showed the need for the RTs when the direction of the overload suddenly shifts. However, a more common case is for the two systems to simply return to normal loading, after which no sharing in either direction is desired. We now show that RTs can cause serious problems when the system returns to normal loading after an overload incident if the activation thresholds are too small. In this case, there is a potential difficulty when the inefficient sharing condition holds, i.e., when $\mu_{1,1} > \mu_{2,1}$ and $\mu_{2,2} > \mu_{1,2}$, which is what we now assume. In this case, RTs combined with small activation thresholds can lead to oscillatory poor performance. Indeed, the small number of shared customers, under the restriction imposed by the RTs, can lead to minor overloads that may trigger undesirable sharing if the activation thresholds are too small. (The simplest example is when $k^n_{1,2} = k^n_{2,1} = 0$, in which case, it is intuitively clear that sharing will be activated and switch sides often.)

We emphasize that, even though the performance is oscillatory, the model after the overload is over a (necessarily aperiodic) positive recurrent and stationary time-homogeneous CTMC when there is abandonment (as discussed in §1), and when the staffing functions and arrival rates are fixed, time-independent functions.

4.1. Simulations of Oscillating Systems with Inefficient Sharing

The oscillatory behavior occurs for systems with abandonment, but it is often hard to detect, because the abandonment ensures that the stationary stochastic system after the overload has ended is stable and it dampens any oscillatory behavior. To demonstrate dramatically, we simulated a system with extreme and unrealistic parameters. (We show a realistic example below.) In this extreme example, we let $\mu_{1,1} = \mu_{2,2} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.1$, and $\theta_1 = \theta_2 = 0.01$. We take ratio parameters $r_{i,j}$, activation thresholds $k^n_{i,j} = 10$, and RTs $\tau^n_{i,j} = 1$ for $i, j = 1, 2$ and $i \neq j$. We start the

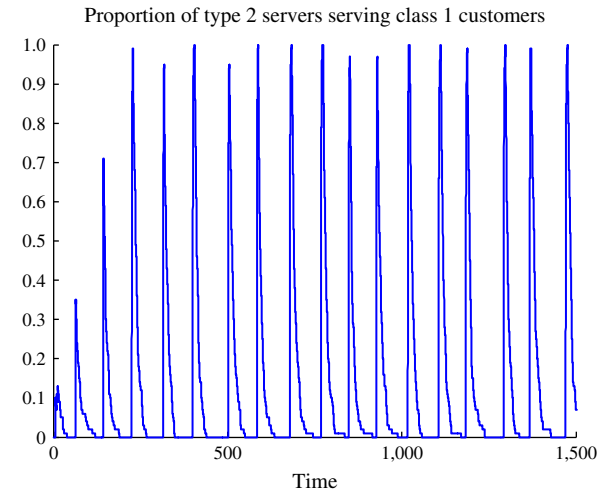

Figure 3 **(Color online) Oscillations of $\bar{Z}^n_{1,2}$ in the Extreme Example**

system with both pools busy serving their own class, but no queues, i.e., $Z^n_{1,1}(0) = Z^n_{2,2}(0) = 100$ and $Q^n_1(0) = Q^n_2(0) = 0$. Figures 3 and 4 show that the oscillatory behavior remains. Moreover, Figure 4 suggests that Q^n_2 (and, by symmetry, also Q^n_1) stabilizes at an overloaded oscillatory equilibrium. The oscillatory behavior in Figures 3–4 may be surprising at first, because the underlying (time-homogeneous) CTMC after the overload has ended is ergodic, as we previously mentioned. Fortunately, the fluid model provides valuable insight, as we explain in §4.2.

We now consider a less extreme, more realistic example in which the sharing service and abandonment rates are changed to $\mu_{1,2} = \mu_{2,1} = \theta_1 = \theta_2 = 0.5$. First, Figure 5 shows the proportion of shared customers over time with the previously specified activation thresholds of $k^n_{i,j} = 10$, but we now consider a system that is recovering from an overload in which

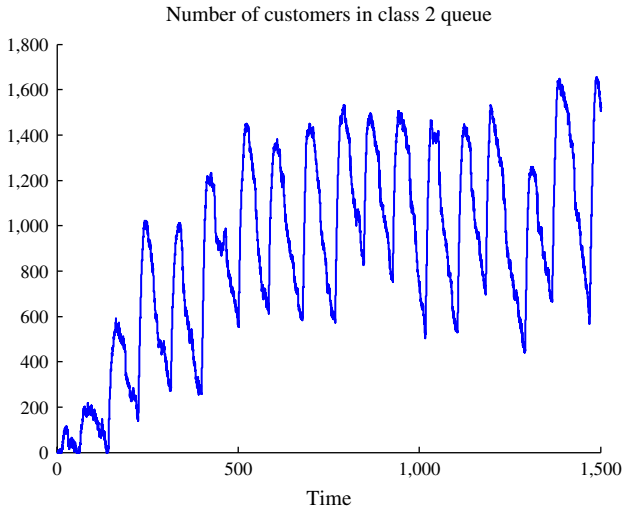

Figure 4 **(Color online) Oscillating Stable Behavior of $\bar{Q}^n_2$ in the Extreme Example**

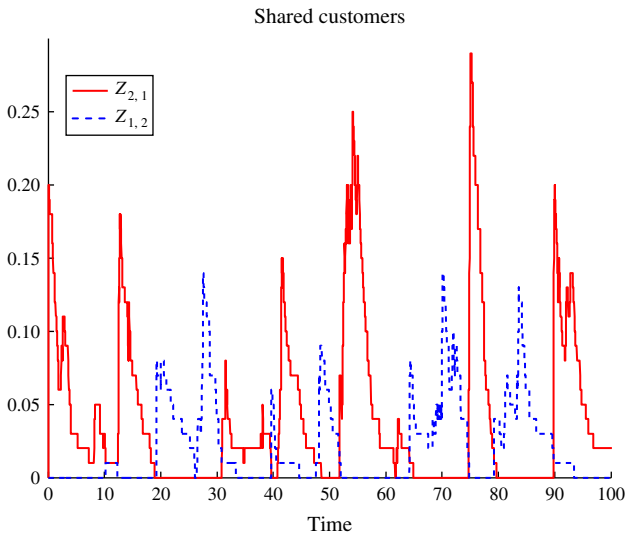

Figure 5 **(Color online) Oscillations of $\bar{Z}^n_{1,2}$ in the More Realistic Example with $k^n_{i,j} = 10$**

pool 1 was helping class-2 customers. In particular, there are initially 20 type-1 agents helping class-2 customers. By taking this initial condition, we are considering a system that starts "worse off" than before, because it is initially overloaded. (In the other two examples, the systems were initialized empty.) We consider the time interval $[0, 100]$ to make the figures clear, but the behavior shown in the figures below remained for the whole duration of the simulation (which lasted for 1,500 time units).

In this case, substantial customer abandonment significantly dampens the sharing oscillations seen previously. Nevertheless, Figure 5 shows that the pools share repeatedly in an oscillating manner over the time interval $[0, 100]$. Although the long-run average number of agents that are helping the other class is not significant, this oscillatory behavior, is clearly undesirable. We do not show figures of the queues because they are uninformative (the oscillations are insignificant). Hence the bad behavior in a system with a relative substantial customer abandonment may be hard to detect by only observing the queues, so that a system with no abandonment or low abandonment rate, gives important insights.

To remedy the problem in Figure 5, we propose increasing the activation thresholds. With larger activation thresholds, the increased stochastic fluctuations due to having some shared customers in both pools do not initiate undesirable sharing. We again refer to Perry and Whitt (2015) for detailed analysis. To illustrate the potential benefit, Figure 6 shows the sharing when the activation thresholds are increased to $k^n_{i,j} = 35$, $i, j = 1, 2$ with all other parameters kept the same. Even though some customers are shared occasionally, especially just after the overload is over, the oscillatory behavior is minimal and decays quickly.

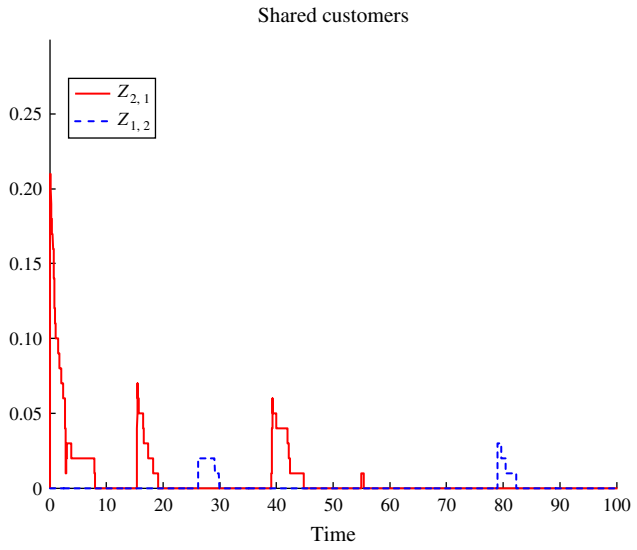

Figure 6 **(Color online) Fewer Oscillations in $\bar{Z}^n_{1,2}$ in the More Realistic Example with $k^n_{i,j} = 35$**

4.2. Insight from the Deterministic Fluid Model

In the examples we have just considered, the six-dimensional stochastic process X^n in (2) describing the system performance after the overload incident has ended is a stationary CTMC. With customer abandonment, that CTMC is necessarily stable, so that with FQR-ART and any parameter setting, the stochastic process X^n in (2) necessarily has a unique steady-state distribution. Nevertheless, we have just seen that the system can exhibit quite complex undesirable behavior for some initial conditions if the control parameters are not set properly.

Fortunately, the fluid model we develop provides an effective means to study the complex system performance and set the control parameters. The oscillating behavior we see in the simulations looks periodic, but it is not quite; it is nearly periodic, like in Liu and Whitt (2011). The system becomes more nearly periodic as the scale increases. In the many-server heavy traffic limit, the stochastic process X^n approaches the deterministic solution of the fluid model we introduce next to serve as an approximation. From the algorithm for that fluid model, we see that it possesses a periodic equilibrium for some initial conditions.

As a consequence, the fluid model can be bistable; it can have a periodic equilibrium in addition to a stable equilibrium, depending on the initial conditions. Consequently, the order in which two different limits occur leads to different stories. As time increases, for any fixed scale, the stochastic process approaches its unique steady-state distribution. In contrast, as the scale increases, a properly scaled version of the stochastic process approaches a deterministic function, which can be periodic. Thus the fluid model provides important insight: an oscillatory fluid approximation implies that a corresponding large system experiences oscillatory behavior for prohibitively large time intervals, even though it is essentially a stationary CTMC. In Perry and Whitt (2015), we prove that the fluid models can exhibit this bistability. Here, it is verified numerically by applying the fluid algorithm.

5. The Fluid Model

The fluid model approximating the stochastic system X^n under FQR-ART is described as the solution to an ordinary differential equation (ODE), but that ODE depends on a stochastic averaging principle (AP). In this section, we derive that ODE via a heuristic representation of the inhomogeneous CTMC in (2). The reasoning in the justification of the fluid model approximation parallels the heuristic engineering discussion in Perry and Whitt (2011a), to which we refer for more discussion. For mathematical support for that reasoning, see Perry and Whitt (2011b, 2013).

5.1. Representation of the Stochastic System During Overloads

The sample paths of the queueing system can be represented in terms of its primitive processes, i.e., the arrival, abandonment, and service processes, as a function of the control. Unlike traditional fluid models, in which the primitive stochastic processes are replaced by their long-run rates, the deterministic fluid model here is more involved and includes a stochastic ingredient in the form of a stochastic AP, which we describe in detail in §5.2.

Even though we are not proving that the fluid model arises as a weak limit of the fluid-scaled stochastic system, we need to take asymptotic considerations to develop the fluid approximation. We thus start with a representation of the stochastic system during overloads, assuming that both service pools are full over an interval $[0, T]$, i.e.,

$$Z^n_{1,1}(t) + Z^n_{2,1}(t) = m^n_1(t) \quad \text{and}$$
$$Z^n_{2,2}(t) + Z^n_{1,2}(t) = m^n_2(t), \quad t \in [0, T]. \tag{10}$$

During the time interval $[0, T]$ no customers can enter service immediately upon arrival, and so all customers are delayed in queue. For simplicity, we first consider intervals over which the staffing functions are continuous and differentiable everywhere. In the online supplement (Figure 13), we give an example of a staffing function with discontinuity.

We represent the sample paths of X^n as random time changes of independent unit rate Poisson processes, as reviewed in Pang et al. (2007). Let

$$\mathcal{A}^n_{1,2}(s) \equiv \left\{\{D^n_{1,2}(s) > 0\} \cap \{Z^n_{2,1}(s) \le \tau^n_{2,1}\}\right\} \quad \text{and}$$
$$\mathcal{A}^n_{2,1}(s) \equiv \left\{\{D^n_{2,1}(s) > 0\} \cap \{Z^n_{1,2}(s) \le \tau^n_{1,2}\}\right\}, \tag{11}$$

the representation of Q_1^n over $[0, T]$ is

$$\begin{aligned} Q_1^n(t) = N_1^a\left(\int_0^t \lambda_1^n(s)\,ds\right) - N_1^u\left(\theta_1\int_0^t Q_1^n(s)\,ds\right) \\ - N_1^+\left(\int_0^t \mathbf{1}_{\mathscr{A}_{1,2}^n(s)}(\mu_{1,1}Z_{1,1}^n(s) + \mu_{1,2}Z_{1,2}^n(s) \right. \\ \left. + \mu_{2,1}Z_{2,1}^n(s) + \mu_{2,2}Z_{2,2}^n(s))\,ds\right) \\ - N_1^-\left(\int_0^t (1 - \mathbf{1}_{\mathscr{A}_{1,2}^n(s)} - \mathbf{1}_{\mathscr{A}_{2,1}^n(s)}) \right. \\ \left. \cdot\,(\mu_{1,1}Z_{1,1}^n(s) + \mu_{2,1}Z_{2,1}^n(s))\,ds\right), \end{aligned}$$

where N_1^a, N_1^u, N_1^+, and N_1^- are mutually independent unit rate (homogeneous) Poisson processes, and $\mathbf{1}_A$ is the indicator function that is equal to 1 if event A occurs, and 0 otherwise.

Note that the representation of Q_1^n is essentially a flow conservation equation (based on the memoryless property of the exponential distribution). That is, the queue at time t is all those customers who arrived by that time, captured by the Poisson process N_1^a, minus all the customers who abandoned, captured by the Poisson process N_1^u, minus all those who were routed into service, as captured by the last two Poisson processes in the expression. Similar expressions hold for the other processes in X^n.

We elaborate on how the intensities of the last two Poisson processes in the right-hand side (RHS) of the representation were obtained. First, if at time $s \in [0, T]$, the event $\mathscr{A}_{1,2}^n(s)$ in (11) holds, then any newly available agent in the system will take his next customer from the head of queue 1. Because agents become available at an instantaneous rate $\sum_{i,j}\mu_{i,j}Z_{i,j}^n(s)$ at time s, we get the third component in the RHS of $Q_1^n(t)$. Next, we recall that, by the routing rule of FQR-ART, if at a time $s \in [0, T]$ $\mathscr{A}_{2,1}^n(s)$ in (11) holds, then any newly available agent takes his next customer from queue 2, in which case queue 1 will not decrease due to a service completion. If neither of the events $\mathscr{A}_{1,2}^n(s)$ or $\mathscr{A}_{2,1}^n(s)$ holds at a time s, then only service completions at pool 1 will cause a decrease at queue 1 due to a customer from that queue being routed to service. That explains the last term in the RHS of the representation.

Next, we exploit the fact that each of the Poisson processes in the representation minus its random intensity constitutes a martingale; again, see Pang et al. (2007) and Perry and Whitt (2013); e.g.,

$$M_1^{n,u} \equiv N_1^u\left(\theta_1\int_0^t Q_1^n(s)\,ds\right) - \theta_1\int_0^t Q_1^n(s)\,ds$$

is a martingale. Thus, subtracting and then adding all the random intensities, and using the fact that a sum of martingales is again a martingale, we get the following representation for the processes Q_1^n, Q_2^n, $Z_{1,2}^n$, $Z_{2,1}^n$ (the remaining two processes $Z_{1,1}^n$ and $Z_{2,2}^n$ are determined by (10)):

$$\begin{aligned} Q_1^n(t) = {} & M_1^n(t) + \int_0^t \lambda_1^n(s)\,ds - \int_0^t \theta_1 Q_1^n(s)\,ds \\ & - \int_0^t \mathbf{1}_{\mathscr{A}_{1,2}^n(s)}(\mu_{1,1}Z_{1,1}^n(s) + \mu_{1,2}Z_{1,2}^n(s) \\ & \qquad + \mu_{2,1}Z_{2,1}^n(s) + \mu_{2,2}Z_{2,2}^n(s))\,ds \\ & - \int_0^t (1 - \mathbf{1}_{\mathscr{A}_{1,2}^n(s)} - \mathbf{1}_{\mathscr{A}_{2,1}^n(s)}) \\ & \qquad \cdot\,(\mu_{1,1}Z_{1,1}^n(s) + \mu_{2,1}Z_{2,1}^n(s))\,ds, \\ Q_2^n(t) = {} & M_2^n(t) + \int_0^t \lambda_2^n(s)\,ds - \int_0^t \theta_2 Q_2^n(s)\,ds \\ & - \int_0^t \mathbf{1}_{\mathscr{A}_{2,1}^n(s)}(\mu_{1,1}Z_{1,1}^n(s) + \mu_{1,2}Z_{1,2}^n(s) \\ & \qquad + \mu_{2,1}Z_{2,1}^n(s) + \mu_{2,2}Z_{2,2}^n(s))\,ds \\ & - \int_0^t (1 - \mathbf{1}_{\mathscr{A}_{1,2}^n(s)} - \mathbf{1}_{\mathscr{A}_{2,1}^n(s)}) \\ & \qquad \cdot\,(\mu_{2,2}Z_{2,2}^n(s) + \mu_{1,2}Z_{1,2}^n(s))\,ds, \\ Z_{1,2}^n(t) = {} & M_{1,2}^n(t) + \int_0^t \mathbf{1}_{\mathscr{A}_{1,2}^n(s)}\mu_{2,2}Z_{2,2}^n(s)\,ds \\ & - \int_0^t (1 - \mathbf{1}_{\mathscr{A}_{1,2}^n(s)})\mu_{1,2}Z_{1,2}^n(s)\,ds, \\ Z_{2,1}^n(t) = {} & M_{2,1}^n(t) + \int_0^t 1_{\mathscr{A}_{2,1}^n(s)}\mu_{1,1}Z_{1,1}^n(s)\,ds \\ & - \int_0^t (1 - \mathbf{1}_{\mathscr{A}_{2,1}^n(s)})\mu_{2,1}Z_{2,1}^n(s)\,ds, \end{aligned} \tag{12}$$

where M_1^n, M_2^n, $M_{1,2}^n$, and $M_{2,1}^n$ are the martingale terms alluded to above. It is not hard to show that those martingales are negligible in the fluid scaling (divided by n, e.g., $\bar{M}_i^n \equiv n^{-1}M_i^n$), i.e., that $\bar{M}_i^n \Rightarrow 0$ and $\bar{M}_{i,j}^n \Rightarrow 0$ as $n \to \infty$, uniformly over $[0, T]$, $i, j = 1, 2$; see, e.g., Perry and Whitt (2013, Lemma 6.1). Hence we consider those martingales like a negligible stochastic noise that can be ignored for the purpose of developing the fluid approximation for (12). The resulting fluid approximation is for the fluid-scaled process $\bar{X}^n \equiv n^{-1}X^n$, applying as n gets large.

When we let $n \to \infty$, we want to replace the stochastic integral representation in (12) with a deterministic one. To do that, we need to replace the indicator functions with smooth functions. This is where the AP comes in. What we do is replace the term $1_{\{D_{1,2}^n(t)>0\}}$ by the steady-state probability that an associated fast-time-scale process (FTSP) is greater than or equal to 0, denoted by $\pi_{1,2}(x(t))$, which is a function of the fluid state at time t, $x(t)$. Both $x(t)$ and $\pi(x(t))$ turn out to be a continuous function of t. This complicated step requires more explanation and justification, which again was the subject of our previous papers.

We give a brief account in the rest of this section and the following one. We start by assuming that there is a fluid limit $\bar{X}^n \Rightarrow x$ as $n \to \infty$ for X^n in (12), where the limit x is a deterministic function that is *continuous and differentiable*. (This fact can be shown to hold by a minor modification of Perry and Whitt 2013, Corollary 5.1.) For any fluid point $x(t)$, let

$$d_{1,2}(x(t)) \equiv q_1(t) - r_{1,2}q_2(t) - k_{1,2} \quad \text{and}$$
$$d_{2,1}(x(t)) \equiv r_{2,1}q_2(t) - q_1(t) - k_{2,1}.$$

We first observe that, if $d_{i,j}(x(t)) > 0$, then since $d_{i,j}(\cdot)$ is a continuous function, $d_{i,j}$ is strictly positive over an interval, and similarly if $d_{i,j} < 0$, $i, j = 1, 2$. In such cases, the indicator functions are easy to deal with because each is a constant over the interval, and equals either 1 or 0. For example, if $d_{1,2}(x(t)) > 0$ for $t \in [s_1, s_2)$ for some $0 \le s_1 < s_2 < \infty$, and in addition, $Z^n_{2,1}(t) \le \tau^n_{2,1}$ over that interval for all n large enough, then

$$\mathbf{1}_{\mathcal{A}^n_{1,2}(t)} \equiv \mathbf{1}_{\{\{D^n_{1,2}(t)>0\}\cap\{Z^n_{2,1}(t)\le\tau^n_{2,1}\}\}} = \mathbf{1}_{[s_2,s_2)}(t), \quad \text{for all } n \text{ large enough.} \tag{13}$$

Hence, a careful study is required for all $x(t) = \gamma$ in the *boundary sets* defined by

$$\mathbb{B}_{1,2} \equiv \{\gamma \in \mathbb{R}_6 \colon d_{1,2}(\gamma) = 0\} \quad \text{and}$$
$$\mathbb{B}_{2,1} \equiv \{\gamma \in \mathbb{R}_6 \colon d_{2,1}(\gamma) = 0\}.$$

FQR-ART aims to "pull" the fluid model to one of these two boundary sets during overloads, when sharing is actively taking place, i.e., $\mathbb{B}_{i,j}$ is the region of the state space where we aim the fluid model to be when pool j helps class i, $i, j = 1, 2$.

Unfortunately, there is no straightforward fluid counterpart to the stochastic processes $D^n_{1,2}$ and $D^n_{2,1}$ when the fluid is in the boundary sets. More specifically, if $x(t) \in \mathbb{B}_{1,2}$ for all t in some time interval $[s_1, s_2)$, then $D^n_{1,2}$ will fluctuate about the threshold $k^n_{1,2}$ over that time interval for all $n \ge 1$. In that case, the indicator function in the RHS of (13) cannot be replaced by 1 or 0 in the fluid model, because the probability that $D^n_{1,2} \ge k^n_{1,2}$ does not converge to 0 or 1 as $n \to \infty$ over $[s_1, s_2)$. The fluid dynamics in the boundary set $\mathbb{B}_{i,j}$ require a refined analysis of the corresponding difference process $D^n_{i,j}$, which takes into account asymptotic reasonings, as we explain later.

5.2. A Stochastic Averaging Principle

For the discussion now, assume that the fluid limit function x is at the boundary set $\mathbb{B}_{1,2}$ over an interval, i.e., $x \in \mathbb{B}_{1,2}$ over some interval $[t_1, t_2]$, and consider the prelimit process $D^n_{1,2}$ over that time interval. As explained in the paragraph above, this implies that $D^n_{1,2}/n \Rightarrow d_{1,2} \equiv \mathbf{0}$ as $n \to \infty$, where $\mathbf{0}$ denotes here the function that is identically 0, so that it is not immediately clear what is the limit of the indicator functions $\mathbf{1}_{\mathcal{A}^n_{i,j}}$, $i, j = 1, 2$. To determine the fluid limit of $\bar{X}^n$ in that case, we must consider the behavior of the unscaled process $D^n_{1,2}$ over $[t_1, t_2]$, paralleling the analysis in Perry and Whitt (2013).

Specifically, to apply the results in Perry and Whitt (2013), we assume (for now) that the arrival rates are fixed (the arrival processes are homogeneous Poisson processes) and that $Z^n_{2,1} < \tau^n_{2,1}$, so that routing is determined solely on the value of $D^n_{1,2}$. In particular, sharing can take place if $D^n_{1,2}(t) > 0$. Then, by Perry and Whitt (2013, Theorem 4.5),

$$D^n_{1,2}(t) \Rightarrow D_{1,2}(x(t), \infty), \quad \text{in } \mathbb{R} \text{ as } n \to \infty, \tag{14}$$

where t is fixed (the convergence holds in $\mathbb{R}$), and $D_{1,2}(\gamma, \cdot) \equiv \{D_{1,2}(\gamma, s) \colon s \ge 0\}$ is a CTMC associated with $\gamma \in \mathbb{R}_6$ whose distribution is determined by the value γ. (There is a different process for each γ.)

An analogous result holds for $D^n_{2,1}$ when the fluid limit satisfies $x \in \mathbb{B}_{2,1}$ over an interval. The notation $D_{i,j}(\gamma, \infty)$ stands for a random variable that has the steady-state distribution of the CTMC $D_{i,j}(\gamma, \cdot)$. Loosely speaking, $D^n_{i,j}$ moves so fast when x is in $\mathbb{B}_{i,j}$, that it reaches its steady state instantaneously as $n \to \infty$. Hence we call the limiting process $D_{i,j}(\gamma, \cdot)$ the FTSP associated with the point γ, or simply the FTSP. Since we are interested in analyzing the indicator functions in (12), we first define for all $\gamma \in \mathbb{R}_6$

$$D_{i,j}(\gamma, \cdot) \equiv +\infty \quad \text{if } d_{i,j}(\gamma) > 0 \quad \text{and}$$
$$D_{i,j}(\gamma, \cdot) \equiv -\infty \quad \text{if } d_{i,j}(\gamma) < 0.$$

Next, for $\gamma \in \mathbb{B}_{1,2}$ and $\gamma \in \mathbb{B}_{2,1}$, respectively, we define

$$\pi_{1,2}(\gamma) \equiv P(D_{1,2}(\gamma, \infty) > 0) \quad \text{and}$$
$$\pi_{2,1}(\gamma) \equiv P(D_{2,1}(\gamma, \infty) > 0). \tag{15}$$

Now, by Perry and Whitt (2013, Theorem 4.1), which was proved for the process $D^n_{1,2}$ when $x \in \mathbb{B}_{1,2}$, and assuming that $Z^n_{2,1}(s) \le \tau^n_{2,1}$ over $[t_1, t_2]$ for all n large enough, we have that, as $n \to \infty$,

$$\int_{t_1}^{t_2} \mathbf{1}_{\mathcal{A}^n_{1,2}(s)}\, ds \equiv \int_{t_1}^{t_2} \mathbf{1}_{\{\{D^n_{1,2}(s)>0\}\cap\{Z^n_{2,1}(s)\le\tau^n_{2,1}\}\}}\, ds \Rightarrow \int_{t_1}^{t_2} \pi_{1,2}(x(s))\, ds.$$

Similarly, if $x \in \mathbb{B}_{2,1}$ over an interval $[t_3, t_4]$, and $Z^n_{1,2}(s) \le \tau^n_{1,2}$ for all n large enough over that interval, we have $\int_{t_3}^{t_4} \mathbf{1}_{\mathcal{A}^n_{2,1}(s)}\, ds \Rightarrow \int_{t_3}^{t_4} \pi_{2,1}(x(s))\, ds$. The convergence in both equations holds uniform. We called these limits a "stochastic averaging principle," or simply an AP, since the process $D^n_{i,j}(t)$ is replaced by the *long-run average* behavior of the corresponding FTSP

$D_{i,j}(x(t),\cdot)$ for each time t over the appropriate interval. If the family of FTSP's $D_{i,j}(x(t),\cdot)$ is positive recurrent for all $t \in I \equiv [t_1, t_2)$, then the AP implies that SSC holds over the time interval I, because the stochastic fluctuations of the FTSP, and therefore of its prelimit $D^n_{i,j}(t)$, $t \in I$, are $o_P(n)$, where $o_P(n)$ denotes a random variable satisfying $o_P(n)/n \Rightarrow 0$ as $n \to \infty$; see Perry and Whitt (2013, Theorem 4.5 and Corollary 4.1).

In the FQR-ART settings, the AP holds under the assumption that $Z^n_{i,j}$ lies below the appropriate RT over the interval $[t_1, t_2]$ for all n large enough (i.e., with probability converging to 1 as $n \to \infty$). If $Z^n_{i,j}$ is larger than the appropriate RT for all n large enough (again, with probability converging to 1) over $[t_1, t_2]$, then the limit of the integral considered above is clearly the 0 function. It remains to rigorously prove convergence theorems at points at which $Z^n_{i,j}(t) = \tau^n_{i,j} + o_P(n)$. However, it is not hard to determine what the dynamics of the limit should be at such points if the limit exists, as shown in the following heuristic fluid approximation.

5.3. Representation via an ODE

The heuristic limiting arguments lead to the following fluid approximation for the X system under FQR-ART during overload periods. Considering an interval $[0, T]$ for which $z_{1,1}(t) + z_{2,1}(t) = m_1(t)$ and $z_{2,2}(t) + z_{2,2}(t) = m_2(t)$ for all $t \in [0, T]$, together with an initial condition $x(0)$, the fluid model of X^n is the solution $x \equiv \{x(t)\colon t \geq 0\}$ over $[0, T]$ to the ODE:

$$
\begin{aligned}
\dot{q}_1(t) &= \lambda_1(t) - \theta_1 q_1(t) - \Pi_{1,2}(x(t)) \\
&\quad \cdot \big(\mu_{1,1} z_{1,1}(t) + \mu_{1,2} z_{1,2}(t) \\
&\qquad + \mu_{2,1} z_{2,1}(t) + \mu_{2,2} z_{2,2}(t)\big) \\
&\quad - \big(1 - \Pi_{1,2}(x(t)) - \Pi_{2,1}(x(t))\big) \\
&\quad \cdot \big(\mu_{1,1} z_{1,1}(t) + \mu_{2,1} z_{2,1}(t)\big), \\
\dot{q}_2(t) &= \lambda_2(t) - \theta_2 q_2(t) - \Pi_{2,1}(x(t)) \\
&\quad \cdot \big(\mu_{1,1} z_{1,1}(t) + \mu_{1,2} z_{1,2}(t) \\
&\qquad + \mu_{2,1} z_{2,1}(t) + \mu_{2,2} z_{2,2}(t)\big) \\
&\quad - \big(1 - \Pi_{1,2}(x(t)) - \Pi_{2,1}(x(t))\big) \\
&\quad \cdot \big(\mu_{2,2} z_{2,2}(t) + \mu_{1,2} z_{1,2}(t)\big), \\
\dot{z}_{1,2}(t) &= \Pi_{1,2}(x(t)) \mu_{2,2} z_{2,2}(t) \\
&\quad - \big(1 - \Pi_{1,2}(x(t))\big) \mu_{1,2} z_{1,2}(t), \\
\dot{z}_{2,1}(t) &= \Pi_{2,1}(x(t)) \mu_{1,1} z_{1,1}(t) \\
&\quad - \big(1 - \Pi_{2,1}(x(t))\big) \mu_{2,1} z_{2,1}(t), \\
\dot{m}_1(t) &= \dot{z}_{1,1}(t) + \dot{z}_{2,1}(t), \\
\dot{m}_2(t) &= \dot{z}_{2,2}(t) + \dot{z}_{1,2}(t),
\end{aligned}
\tag{16}
$$

where, for $\pi_{i,j}(x(t))$ in (15), $i, j = 1, 2$,

$$
\Pi_{i,j}(x(t)) := \begin{cases} \pi_{i,j}(x(t)), & \text{if } z_{j,i}(t) \leq \tau_{j,i}; \\ 0, & \text{otherwise.} \end{cases}
$$

We remark that the ODE (16) can be equivalently represented by an integral equation resembling (12), but with the negligible martingale terms omitted, all the stochastic processes replaced by their fluid counterparts, and the indicator functions replaced by the appropriate $\Pi_{i,j}$ functions.

In practice, we do not a priori know the value of T, and there is a need to make sure that the ODE is a valid approximation for the stochastic system. We consider the ODE (16) valid (i.e., a legitimate representation of the evolution of the system) as long as the following two conditions are satisfied: (i) the two queues are strictly positive and (ii) if a queue is equal to 0 at some time $t \geq 0$, then the derivative of that queue is nonnegative at time t (so that the queue is nondecreasing at this time). When the ODE (16) is not valid, then other fluid models should be employed to approximate the system. We discuss such scenarios in the online supplement.

We elaborate on condition (ii). Consider, for example, the ODE for q_1 and assume that $q_1(t) = 0$ and $\dot{q}_1(t) < 0$ for some $t \geq 0$. Necessarily $\Pi_{1,2}(x(t)) = 0$, because $d_{1,2}(x(t)) \leq 0$, and the assumption that $\dot{q}_1(t) < 0$ implies that

$$
\lambda_1(t) - (1 - \Pi_{2,1}(x(t)))(\mu_{1,1} z_{1,1}(t) + \mu_{2,1} z_{2,1}(t)) < 0. \tag{17}
$$

In addition, since all the class-1 arrivals must immediately enter service (for otherwise, the queue will be increasing), it also holds that $\dot{z}_{1,1}(t) = \lambda_1(t) - \mu_{1,1} z_{1,1}(t)$. Hence

$$
\begin{aligned}
&\dot{z}_{1,1}(t) + \dot{z}_{2,1}(t) \\
&= \lambda_1(t) - \mu_{1,1} z_{1,1}(t) + \Pi_{2,1}(x(t)) \mu_{1,1} z_{1,1}(t) \\
&\quad - \big(1 - \Pi_{2,1}(x(t))\big) \mu_{2,1} z_{2,1}(t) \\
&= \lambda_1(t) - \big(1 - \Pi_{2,1}(x(t))\big)\big(\mu_{1,1} z_{1,1}(t) + \mu_{2,1} z_{2,1}(t)\big)
\end{aligned}
\tag{18}
$$

so that, by (17), $\dot{z}_{1,1}(t) + \dot{z}_{2,1}(t) < 0$.

Now, since $\dot{m}_1(t) = \dot{z}_{1,1}(t) + \dot{z}_{2,1}(t)$, we see that pool 1 can remain full just after time t only if $m_1(t)$ happens to decrease exactly as in (18). However, q_1 is becoming negative, so that the ODE is not valid. In contrast, if (18) holds (which ODE (16) enforces to be equal to $\dot{m}_1(t)$) and $q_1(t) = 0$, then necessarily $\dot{q}_1(t) < 0$, so that the queue is becoming negative. In either case, we see that the ODE is valid as an approximation for the stochastic system when $q_1(t) = 0$ only if pool 1 can be kept full without enforcing q_1 to become negative. Similar reasonings hold for the q_2 and m_2 processes.

Remark 1. A proof of existence of a unique solution to the ODE (16) following the lines of Perry and Whitt (2011b) requires showing that the RHS is a local Lipschitz continuous function of x and is piecewise continuous in t. We do not prove such a result here, but it is important to consider arrival rates and staffing functions that ensure that the right side of the ODE satisfies the piecewise continuity condition in the time argument.

6. Solving the ODE

To appreciate that the algorithm cannot be a routine solution of an ODE, observe that computing the solution to (16) requires computing the two steady-state probabilities, $\pi_{1,2}(x(t))$ and $\pi_{2,1}(x(t))$, for all times t and states $x(t) \in \mathbb{R}_6$. Simplification is achieved when $r_{1,2} = r_{2,1} = 1$, because the FTSP's $D_{i,j}(x(t), \cdot)$, $i, j = 1, 2$, become simple *birth-and-death* (BD) processes. To facilitate the discussion, we thus consider this simpler case and refer to Perry and Whitt (2011b, §6.2) for the treatment of the FTSP $D_{1,2}$ as a *quasi-birth-and-death process* (QBD) when the ratio parameters are not equal to 1. See also Remark 3 and the online supplement.

For simplicity, we again start by assuming that the arrival processes are homogeneous Poisson processes, having constant arrival rates λ_1 and λ_2 over $[0, T]$, and that the staffing functions are also fixed over that time interval at m_1 and m_2. Recall that $D_{i,j}(\gamma, \cdot) \equiv \infty$ if $d_{i,j}(\gamma) > 0$ and $D_{i,j}(\gamma, \cdot) \equiv -\infty$ if $d_{i,j}(\gamma) < 0$, and let $\mathbb{A}_{1,2}$ and $\mathbb{A}_{2,1}$ be the subsets of $\mathbb{R}_6$ in which the FTSP's $D_{1,2}(\gamma, \cdot)$ and $D_{2,1}(\gamma, \cdot)$ are positive recurrent, i.e.,

$$\mathbb{A}_{1,2} \equiv \{\gamma \in \mathbb{B}_{1,2}\colon 0 < \pi_{1,2}(\gamma) < 1\} \quad \text{and}$$
$$\mathbb{A}_{2,1} \equiv \{\gamma \in \mathbb{B}_{2,1}\colon 0 < \pi_{2,1}(\gamma) < 1\}. \tag{19}$$

By definition, if the fluid model at time t is in $\mathbb{A}_{i,j}$, i.e., $x(t) \in \mathbb{A}_{i,j}$, then $d_{i,j}(x(t)) = 0$. However, if $d_{i,j}(x(t)) = 0$, then $x(t)$ is not necessarily in $\mathbb{A}_{i,j}$, because the FTSP $D_{i,j}(x(t), \cdot)$ may be transient (drift to $+\infty$ or $-\infty$) or null recurrent; in particular, the evolution of the fluid model is determined by the distributional characteristics of the FTSPs $D_{1,2}$ and $D_{2,1}$. Hence, even before we try to compute $\pi_{i,j}(x(t))$, which is necessary to solve the ODE (16), there is a need to determine whether $x(t)$ is in one of the sets $\mathbb{A}_{1,2}$ or $\mathbb{A}_{2,1}$. We focus on $D_{1,2}$ with the analysis of $D_{2,1}$ being similar.

To determine the behavior of the FTSP $D_{1,2}$, it is again helpful to think of x as a fluid limit of the fluid-scaled sequence $\{\bar{X}^n\colon n \geq 1\}$ and to recall that $D_{1,2}$ was achieved from $D^n_{1,2}$ in the limit without any scaling; see (14). (See also Perry and Whitt 2013, Theorem 4, which provides a process-level limit relating $D_{1,2}$ and $D^n_{1,2}$.) Hence, both processes are defined on the same state space, which for $r_{1,2} = 1$, is $\mathbb{Z} \equiv \{\ldots, -1, 0, 1, \ldots\}$.

Let $s \geq 0$ denote the fast-time scale, so that for each fixed $x(t)$, $\{D_{1,2}(x(t), s)\colon s \geq 0\}$ is a BD process. The BD rates of the FTSP associated with the point $x(t)$, $D_{1,2}(x(t), \cdot)$, can be inferred from the value of $x(t)$ and the instantaneous random rates of the prelimit process $D^n_{1,2}$: If at a time $s \geq 0$ $D_{1,2}(x(t), s) = m > 0$, the BD rates of the FTSP are, respectively,

$$\lambda^+(x(t), m) \equiv \lambda_1 + \theta_2 q_2(t),$$
$$\mu^+(x(t), m) \equiv \lambda_2 + \mu_{1,1} z_{1,1}(t) + \mu_{1,2} z_{1,2}(t) + \mu_{2,1} z_{2,1}(t) + \mu_{2,2} z_{2,2}(t) + \theta_1 q_1(t).$$

In analogy to the (non-Markov) process $D^n_{1,2} = Q^n_1 - Q^n_2 - k^n_{1,2}$, $\lambda_+(x(t), m)$ corresponds to an increase of $D_{1,2}$ due to arrival to queue 1 plus an abandonment from queue 2 (since either one of these two events cause an increase by 1 of $D^n_{1,2}$ in the stochastic system). Because any other event causes $D^n_{1,2}$ to decrease by 1, due to the scheduling rules of FQR-ART, we get the expression for $\mu^+(x(t), m)$.

Next, if at time $s \geq 0$, $D_{1,2}(x(t), s) = m \leq 0$, the BD rates are, respectively,

$$\lambda^-(x(t), m) \equiv \lambda_1 + \mu_{2,2} z_{2,2}(t) + \mu_{1,2} z_{1,2}(t) + \theta_2 q_2(t),$$
$$\mu^-(x(t), m) \equiv \lambda_2 + \mu_{1,1} z_{1,1}(t) + \mu_{2,1} z_{2,1}(t) + \theta_1 q_1(t).$$

Again, whenever $D^n_{1,2}$ is nonpositive and sharing is taking place with pool 2 helping class 1, a "birth" occurs if there is an arrival to queue 1 or an abandonment from queue 2, or if there is a service completion in pool 2 (since then, a newly available type-2 agent takes his next customer from queue 2). Similarly, a "death" occurs if there is an arrival to class 2, an abandonment from queue 1, or a service completion in pool 1.

We see that the FTSP $D_{1,2}(x(t), \cdot)$ is a two-sided $M/M/1$ queue, i.e., it behaves like an $M/M/1$ queue with "arrival rate" $\lambda^+(x(t), m)$ and "service rate" $\mu^+(x(t), m)$ for all $m > 0$, and behaves like a different $M/M/1$ queue with "arrival rate" $\mu^-(x(t), m)$ and "service rate" $\lambda^-(x(t), m)$ for all $m \leq 0$. Thus, for

$$\delta^+(\gamma) \equiv \lambda^+(\gamma, \cdot) - \mu^+(\gamma, \cdot) \quad \text{and}$$
$$\delta^-(\gamma) \equiv \lambda^-(\gamma, \cdot) - \mu^-(\gamma, \cdot), \quad \gamma \in \mathbb{B}_{1,2},$$

the set $\mathbb{A}_{1,2}$ can be characterized via $\mathbb{A}_{1,2} \equiv \{\gamma \in \mathbb{B}_{1,2}\colon \delta^+(\gamma) < 0 < \delta^-(\gamma)\}$. Next, letting $T^+(\gamma)$ and $T^-(\gamma)$ denote, respectively, the busy period of the $M/M/1$ in the positive region and the busy period of the $M/M/1$ in the negative region, and using simple alternating renewal arguments for the renewal process $D_{1,2}(\gamma, \cdot)$, we have

$$\pi_{1,2}(\gamma) = \frac{E[T^+(\gamma)]}{E[T^+(\gamma)] + E[T^-(\gamma)]}, \tag{20}$$

where from basic $M/M/1$ theory, $E[T^{\pm}(\gamma)] = 1/(\mu^{\pm}(\gamma)-\lambda^{\pm}(\gamma))$. Note that if $d_{1,2}(\gamma)=0$ but $\gamma \notin \mathbb{A}_{1,2}$, then $\pi_{1,2}(\gamma)$ is equal to either 1 or 0. In particular,

$$\begin{aligned} &\text{if } \delta^{+}(\gamma)\geq 0, \quad \text{then } \pi_{1,2}(\gamma)=1 \quad \text{and} \\ &\text{if } \delta^{-}(\gamma)\leq 0, \quad \text{then } \pi_{1,2}(\gamma)=0. \end{aligned} \tag{21}$$

There are no other options, because for any $\gamma = x(t)$ for which both pools are full (as is required for the ODE (16) to be valid), it holds that

$$\delta^{-}(x(t))-\delta^{+}(x(t)) = 2\big(\mu_{1,2}z_{1,2}(t)+\mu_{2,2}z_{2,2}(t)\big) > 0,$$

where the inequality above follows from the fact that $z_{1,2}(t)+z_{2,2}(t)=m_2(t)>0$.

We see that the sets $\mathbb{A}_{i,j}$ and the computation of $\pi_{i,j}(\cdot)$ are completely determined by the staffing, arrival rates, service, and abandonment rates for any given point $\gamma \in \mathbb{R}_6$, where the only points that require careful analysis are those in one of the two sets $\mathbb{B}_{i,j}$. However, recall that we have assumed for simplicity that the arrival rates and staffing functions are not time dependent. If, instead, the arrival rates or the staffing functions are time dependent, then the distribution of the FTSP $D_{i,j}(x(t),\cdot)$ is also time dependent. In particular, given a $\gamma \in \mathbb{R}_6$, we cannot determine whether $D_{1,2}(\gamma,\cdot)$ is positive recurrent, because that may depend on the time $t \in [0,T]$. Thus the sets at which the FTSPs are ergodic are themselves time dependent. Hence, for a full analysis, we would need to consider sets of the form $\{\mathbb{A}_{i,j}(t)\colon t\in[0,T]\}$, where

$$\mathbb{A}_{i,j}(t) \equiv \big\{(\gamma,t)\in \mathbb{B}_{i,j}\times\mathbb{R}_{+}\colon \delta^{+}(\gamma,t)<0<\delta^{-}(\gamma,t)\big\}, \tag{22}$$

where $\delta^{+}(\gamma,t)$ and $\delta^{-}(\gamma,t)$ are the drifts of the FTSP $D_{1,2}(\gamma,\cdot)$ at the point γ at time t. Fortunately, for the purpose of solving the ODE, we do not actually need to characterize the sets $\{\mathbb{A}_{i,j}(t)\colon t\in[0,T]\}$, because we can determine whether $D_{i,j}(x(t),\cdot)$ is ergodic at each time t as we solve the ODE.

6.1. A Numerical Algorithm to Solve the ODE

Given the ODE in (16) with a fully specified RHS at each t, we compute the solution x over an interval $[0,T]$ by employing the classical Euler method, combined with the AP. Given a step size h and the time T, the number of iterations needed is $N \equiv T/h$. Let $\dot{x}=\Psi(x)$, where $\Psi(x)$ is the RHS of the appropriate ODE, e.g., if both pools are full, then $\Psi(x)$ is the RHS of (16). Given $x(0)$, we can compute $x(h)$ using the first Euler step: $x(h)=x(0)+h\Psi(x(0))$. Given $x(h)$, we can compute $\Pi_{1,2}(x(h))$ and $\Pi_{2,1}(x(h))$, if needed, and then compute $x(2h)$ using the second Euler step. In general, the solution to the ODE is computed via

$$x((k+1)h) = x(kh)+h\Psi(x(kh)), \quad 0\leq k\leq N,$$

where at each step, if $x(kh)\in\mathbb{B}_{1,2}$ or $x(kh)\in\mathbb{B}_{2,1}$, we can compute $\Pi_{1,2}(kh)$ and $\Pi_{2,1}(kh)$ as explained earlier.

The algorithm just described remains unchanged when the ratio parameters are general (not equal to 1), except that the sets $\mathbb{A}_{i,j}$ and the computations of $\pi_{i,j}$ are more complicated (the FTSPs are no longer BD processes). We refer to Perry and Whitt (2011b) for these more complicated settings.

To evaluate the RHS in each step, we use the analysis in §6, starting at a given initial condition $x(0)$, since we can now determine the value of $\Pi_{i,j}(x(t))$ for each $t\geq 0$. For example, if at a time $t\geq 0$ $d_{1,2}(x(t))=0$, then we check whether (22) holds, so that $x(t)\in\mathbb{A}_{1,2}(t)$. If $z_{2,1}(t)\leq\tau_{2,1}$, then $\Pi_{1,2}(x(t))=\pi_{1,2}(x(t))$ and it can be computed using (20). If $z_{2,1}(t)>\tau_{2,1}$, then $\Pi_{2,1}(t)=0$. If $d_{1,2}(x(t))=0$ but $x(t)\notin\mathbb{A}_{1,2}(t)$, i.e., if (22) does not hold, then we can determine the value of $\pi_{1,2}(x(t))$, and thus of $\Pi_{1,2}(x(t))$, by computing the drifts of the FTSP and employing (21) (replacing the drifts in (21) with the time-dependent drifts as in (22)). Similarly, we can compute the value of $\Pi_{2,1}(x(t))$ whenever $d_{2,1}(x(t))=0$.

In all other regions of the state space for which both pools are full, i.e., $z_{i,j}(t)+z_{j,i}(t)=m_j(t)$, $i\neq j$, we can easily determine the value of $\pi_{1,2}(x(t))$ by considering whether $d_{i,j}(x(t))$ is bigger or smaller than 0. For example, if at time $t\geq 0$ $d_{1,2}(x(t))>0$, then $\pi_{1,2}(x(t))=1$ and if $d_{1,2}(x(t))<0$, then $\pi_{1,2}(x(t))=0$. This, together with the value of $z_{2,1}(t)$, immediately gives the value of $\Pi_{1,2}(x(t))$.

We need to use other fluid equations when at least one of the two pools is not full. If, for example, $z_{1,1}(t)+z_{2,1}(t)<m_1(t)$, then necessarily $q_1(t)=0<k_{1,2}$, so that $\dot{z}_{1,2}(t)=-\mu_{1,2}z_{1,2}(t)$ and $\dot{z}_{1,1}(t)=\lambda_1(t)-\mu_{1,1}z_{1,1}(t)$. The evolution of $z_{2,1}$ in this case is determined by whether $q_2(t)<k_{2,1}$ or $q_2(t)\geq k_{2,1}$. In the first case, $z_{2,1}(t)$ must be strictly decreasing at time t if it is positive, or remain at 0 otherwise. In the latter case, when $q_2(t)\geq k_{2,1}$, the excess fluid—that is not routed to pool 2 and does not abandon, if such excess fluid exists—is flowing to pool 1. We thus have $\dot{z}_{2,1}(t)$ is equal to

$$\begin{aligned} &-\mu_{2,1}z_{2,1}(t), \quad \text{if } q_2(t)<k_{2,1}, \\ &-\mu_{2,1}z_{2,1}(t)+\big(\lambda_2(t)-\mu_{2,2}z_{2,2}(t)-\mu_{1,2}z_{1,2}(t)-\theta_2 k_{2,1}\big)^{+} \\ &\qquad\qquad \text{if } q_2(t)\geq k_{2,1}. \end{aligned}$$

Similar reasonings lead to the fluid model of $z_{1,2}$ when pool 1 is full, but pool 2 has spare capacity.

If both pools have spare capacity at time t, then $q_1(t)=q_2(t)=0$ and $\dot{z}_{i,j}(t)=-\mu_{i,j}z_{i,j}(t)$ and $\dot{z}_{i,i}(t)=\lambda_i-\mu_{i,i}z_{i,i}(t)$, $i,j=1,2$, $i\neq j$.

Remark 2. If at iteration $k\geq 0$, the solution lies outside the set $\mathbb{B}_{1,2}\cup\mathbb{B}_{2,1}$, then due to the discreteness

of the algorithm, there is a need to ensure that the boundary is not missed in the following iterations. Hence, if in the kth iteration $d_{1,2}(x(kh)) > 0$ (< 0) and in the $(k+1)$st iteration $d_{1,2}((x(k+1)h)) < 0$ (> 0), then the boundary $d_{1,2}$ necessarily was missed, because the fluid is continuous, thereforewe set $d_{1,2}((x(k+1)h)) = 0$. We then check whether $x((k+1)h) \in \mathbb{A}_{1,2}((k+1)h)$, compute $\pi_{1,2}(x(k+1)h)$ and use its value to compute the value in the $(k+2)$nd iteration. It is significant that we do not force the solution to be on the boundary, e.g., we do not compute $q_1((k+1)h)$ and use its value to compute $q_2((k+1)h)$ via

$$q_2((k+1)h) = q_1((k+1)h) - k_{1,2}. \tag{23}$$

We solve the six-dimensional ODE in (16), and if indeed (23) holds whenever it should, then we have a good indication that the algorithm works. That is, we can check at which iteration the boundary $\mathbb{B}_{1,2}$ was hit, and then observe if $q_1(t) - q_2(t) = k_{1,2}$ over an interval for which we have indication that this should hold. (Of course, the solution to the algorithm might leave the boundary for legitimate reasons, i.e., because the fluid model leaves it.)

Remark 3. When $r_{i,j} \neq 1$, the FTSP $D_{i,j}(\gamma, \cdot)$, $\gamma \in \mathbb{B}_{i,j}$, can be represented as a QBD. The only difference in the algorithm is that $\pi_{i,j}(\gamma)$ does not have an explicit representation, as in (20). In that case, we use matrix-geometric methods in each iteration of the algorithm to numerically compute the value of $\pi_{i,j}$ in that iteration. See the online supplement for an elaboration.

7. Conclusions

In this paper, we studied a time-varying X model experiencing periods of overloads. Although our previous FQR-T control is effective in automatically responding quickly to unexpected overloads, the examples in §§3 and 4 show that it needs to be modified to recover rapidly after the overload is over, due to either a return to normal loading or a sudden change in the direction of the overload. We thus proposed the FQR-ART control. With FQR-ART, the one-way sharing rule is relaxed by adding the lower RTs. To avoid oscillations of the service process, which, in turn, can cause congestion collapse, we indicated that the activation thresholds also need to be increased, being asymptotically of order $O(n)$ as in (6) instead of $o(n)$, as in (5) with FQR-T.

We then extended the fluid model developed in Perry and Whitt (2011a, b, 2013) based on the stochastic averaging principle to cover a more general time-varying environment, and developed the corresponding algorithm to numerically compute the performance functions in that fluid model. Simulation experiments indicate that this fluid model captures the main dynamics of the system, even in very extreme cases; see the online supplement. Thus the fluid model can be used to ensure that the control parameters of FQR-ART are properly set .

There are many directions for future research. First, it remains to investigate the performance of FQR-ART in more complex time-varying scenarios. Second, it remains to establish theoretical properties of the new fluid model, paralleling Perry and Whitt (2011b). Third, it remains to establish many-server heavy-traffic limits in this more general setting, paralleling Perry and Whitt (2013, 2014).

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/ijoc.2015.0642.

Acknowledgments

The first author received support from National Science Foundation (NSF) [Grant CMMI 1436518]. The second author received support from NSF [Grants CMMI 1066372 and 1265070].

References

Aksin Z, Armony M, Mehrotra V (2007) The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* 16(6):655–688.

Erramilli A, Forys LJ (1991) Oscillations and chaos in a flow model of a switching system. *IEEE J. Selected Areas Comm.* 9(2): 171–178.

Feinberg EA, Reiman MI (1994) Optimality of randomized trunk reservation. *Probab. Engrg. Inform. Sci.* 8(4):463–489.

Feldman Z, Mandelbaum A, Massey WA, Whitt W (2008) Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* 54(2):324–338.

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.

Gurvich I, Whitt W (2009a) Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.* 34(2):363–396.

Gurvich I, Whitt W (2009b) Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* 11(2):237–253.

Gurvich I, Whitt W (2010) Service-level differentiation in many-server service systems via queue-ratio routing. *Oper. Res.* 58(2): 316–328.

Kelly FP (1991) Loss networks. *Ann. Appl. Probab.* 1(3):319–378.

Liu Y, Whitt W (2011) Nearly periodic behavior in the overloaded G/D/S+GI queue. *Stochastic Systems* 1(2):340–410.

Liu Y, Whitt W (2012a) The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems* 71(4):405–444.

Liu Y, Whitt W (2012b) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* 60(6): 1551–1564.

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

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

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

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

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

Perry O, Whitt W (2014) Diffusion approximation for an overloaded X model via a stochastic averaging principle. *Queueing Systems* 76(4):347–401.

Perry O, Whitt W (2015) Chattering and congestion collapse in an overload switching control. Working paper, Columbia University. http://www.columbia.edu/~ww2040/Periodic042315.pdf.

Shah D, Wischik D (2011) Fluid models of congestion collapse in overloaded switched networks. *Queueing Systems* 69(2): 121–143.

Weber JH (1964) A simulation study of routing control in communication networks. *Bell System Tech. J.* 43(6):2639–2676.

Wilkinson RI (1956) Theory for toll traffic engineering in the U.S.A. *Bell System Tech. J.* 35(2):421–513.

Young RM (1991) Euler's constant. *Math. Gazette* 75(472):187–190.