OPERATIONS RESEARCH
Vol. 59, No. 5, September–October 2011, pp. 1159–1170
ISSN 0030-364X | EISSN 1526-5463 | 11 | 5905 | 1159

http://dx.doi.org/10.1287/opre.1110.0985

A Fluid Approximation for Service Systems Responding to Unexpected Overloads

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

In a recent paper we considered two networked service systems, each having its own customers and designated service pool with many agents, where all agents are able to serve the other customers, although they may do so inefficiently. Usually the agents should serve only their own customers, but we want an automatic control that activates serving some of the other customers when an unexpected overload occurs. Assuming that the identity of the class that will experience the overload or the timing and extent of the overload are unknown, we proposed a queue-ratio control with thresholds: When a weighted difference of the queue lengths crosses a prespecified threshold, with the weight and the threshold depending on the class to be helped, serving the other customers is activated so that a certain queue ratio is maintained. We then developed a simple deterministic steady-state fluid approximation, based on flow balance, under which this control was shown to be optimal, and we showed how to calculate the control parameters. In this sequel we focus on the fluid approximation itself and describe its transient behavior, which depends on a heavy-traffic averaging principle. The new fluid model developed here is an ordinary differential equation driven by the instantaneous steady-state probabilities of a fast-time-scale stochastic process. The averaging principle also provides the basis for an effective Gaussian approximation for the steady-state queue lengths. Effectiveness of the approximations is confirmed by simulation experiments.

Subject classifications: large-scale service systems; overload control; many-server queues; fluid approximation; averaging principle; separation of time scales; differential equation; heavy traffic.
Area of review: Stochastic Models.
History: Received September 2008; revisions received July 2009, July 2010, December 2010; accepted January 2011.

1. Introduction

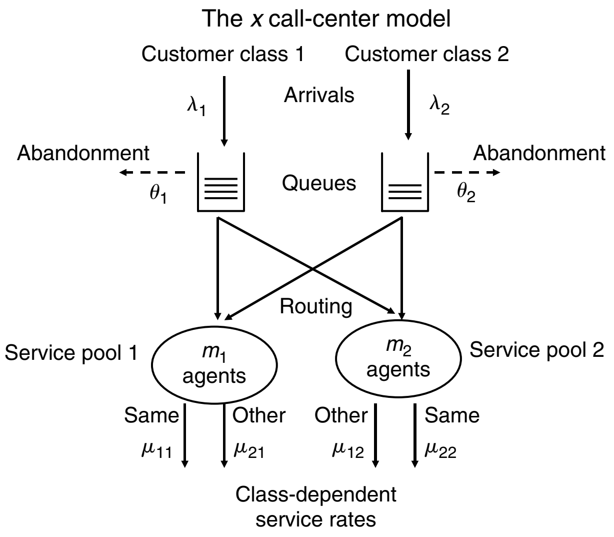
Responding to unexpected overloads. In Perry and Whitt (2009), we considered how two service systems that normally operate independently, such as call centers, can help each other when one encounters an unexpected overload and is unable to immediately increase its own staffing. We assumed that each service system has a service pool with many agents, each of whom has the ability to serve customers from the other system as well as its own, even though the other customers may be served inefficiently. The goal was to find a way to automatically respond to overloads, without knowledge of the arrival rates, while producing only negligible sharing under normal loads.

Toward that end, we proposed a queue-ratio control with thresholds (QR-T) that activates serving customers from the other system when a weighted difference of the two queue lengths exceeds a threshold, allowing sharing in only one direction at any time. There is a target queue-ratio function and threshold for each direction of sharing. The general QR-T control allows the queue ratio to be a function of the two queue lengths, but it often suffices to use fixed queue ratios (FQR-T), which is advantageous because the control then has fewer parameters, namely, the two ratio parameters and the two thresholds.

These queue-ratio controls are modifications of ones proposed previously in Gurvich and Whitt (2009a, b, 2010). The thresholds and the application to respond to unexpected overloads are new. These QR controls tend to be effective because they simplify the problem by reducing the dimension. With the QR-T control, the two system queues tend to evolve independently when the sharing is not activated (under normal loads), but the two queues tend to evolve together in a fixed relation when the sharing is activated (under overloads). Indeed, under overloads, the two system queues tend to evolve dependently to the maximum extent. This maximum dependence can be formalized by the notion of *state-space collapse* (SSC), as in Bramson (1998), Dai and Tezcan (2011), and Gurvich and Whitt (2009a).

The Markovian X **model.** To analyze this QR-T overload control and determine appropriate control parameters, we considered a Markovian X call-center model, as depicted in Figure 1, having two customer classes, each with its own queue, and two service pools, each with many agents; see Aksin et al. (2007), Gans et al. (2003), and

Figure 1. The X model.



Garnett and Mandelbaum (2000) for background on the basic call-center models.

A fluid model with convex costs. In order to determine appropriate queue-ratio functions and to approximate the performance of this QR-T control, we introduced a convex-cost framework and a simple deterministic steady-state fluid approximation for the X call-center model. Within that framework, we proved that properly chosen queue-ratio functions minimize the average steady-state cost during an overload incident without requiring knowledge of the arrival rates. Moreover, we showed how to calculate the optimal queue-ratio functions. In addition, we indicated how to determine the thresholds. We then applied simulation to show that the optimal control for the fluid model is effective for the original stochastic X call-center model. Indeed, the simulations show that the proposed queue-ratio control with thresholds outperforms the optimal fixed partition of the servers given known fixed arrival rates during the overload, even though the proposed control does not use information about the arrival rates.

The contributions here. The present paper develops an approximation for the stochastic processes describing the performance of the overloaded X model with the ratio control. The approximation is interesting because it involves a heavy-traffic *averaging principle* (AP). First, the AP directly yields an approximation for the transient behavior as well as the steady-state behavior. The new approximation for the transient behavior is a deterministic fluid approximation, i.e., an *ordinary differential equation* (ODE), but it is an unconventional ODE. As a consequence of the AP, the ODE is driven by a function of the ODE state involving the steady-state probability distributions of an associated family of fast-time-scale stochastic processes; see §3. The most familiar example of an AP is no doubt in the theory of nearly completely decomposable (NCD) Markov chains, as in Courtois (1977); see Remark 2.4.1 of Whitt (2002) for more discussion and references. We validated the transient approximation based on the ODE by conducting simulation experiments; see §3.

We also apply the AP to develop improved approximations for the steady-state distribution. The heuristic steady-state fluid approximation developed in Perry and Whitt (2009) provides only an approximation for the mean queue lengths. First, the AP provides improved approximations for these mean steady-state values; see §5. Effectiveness is confirmed by simulations in §6. Second, the AP provides a tractable approximation for the full steady-state joint distribution of the queue lengths during the overload incident. In particular, the AP leads to a Gaussian approximation, with explicit formulas for the variances; see §7. The full distribution provided by this Gaussian approximation is vital because, for typical system sizes, the standard deviations tend to be of roughly the same order as the mean values.

The many-server heavy-traffic regime. The performance of the X model during overloads, including the AP, can be understood by considering the *many-server heavy-traffic* (MSHT) limiting regime, briefly reviewed here in §2. Based on an understanding of the MSHT *efficiency-driven* (ED) regime, we see that an AP is appropriate here. We justify these approximations empirically through extensive simulation experiments.

In subsequent papers, Perry and Whitt (2011a, b, c), we put the AP and the associated performance approximations on a firm mathematical basis. We show that the ODE stemming from the AP is well defined, with good properties, and that the approximations we develop in this paper arise as MSHT stochastic process limits involving the AP, paralleling earlier papers by Hunt and Kurtz (1994), and Coffman et al. (1995). We contribute here by showing how the SSC and the AP associated with the MSHT regime can be applied directly as engineering principles.

Even though we do not do any proofs here, we do verify the accuracy of our approximations empirically with simulation. We demonstrate the convergence as $n \to \infty$ in the MSHT limit by showing the performance of the scaled processes for several values of n, in particular, for $n = 25$, 100, and 400. We see remarkable accuracy for $n = 400$ and surprisingly good rough approximations even for $n = 25$. We also see the rapid convergence to steady state as $t \to \infty$. Additional material appears in a longer version maintained on the authors' webpages.

2. Preliminaries

The model. The Markovian X model is depicted in Figure 1. There are two customer classes, with customers from each class arriving according to a Poisson process. There is a queue for each customer class, from which customers are served in order of arrival. Waiting customers have limited patience: A class-i customer will abandon if he does not start service before a random time that is exponentially distributed with mean $1/\theta_i$. There are two service pools, with pool j having m_j homogeneous servers working in parallel. The mean service time for a class-i customer served by a type-j agent is $1/\mu_{i,j}$, which may depend on both the customer class i and the service pool j. The service times, abandonment times, and arrival processes are mutually independent. Let $Q_i(t)$ be the number of class-i customers in queue and let $Z_{i,j}(t)$ be the number of type-j agents busy serving class-i customers at time t. With the assumptions above, the stochastic process $\{(Q_i(t), Z_{i,j}(t); i=1,2; j=1,2)\colon t \geqslant 0\}$ is a six-dimensional continuous-time Markov chain (CTMC), given any routing policy that depends on the six-dimensional state.

We are using this model to describe the system during the overload incident. Our approximation applies after the arrival rates have shifted to new values and after sharing has begun. We assume that customers from the two classes arrive during the overload with constant arrival rates λ_1 and λ_2, which make at least one class overloaded. Our goal is to develop approximations for the stochastic process $\{(Q_i(t), Z_{i,j}(t); i=1,2; j=1,2)\colon t \geqslant 0\}$ during the overload incident.

The FQR-T control. The FQR-T control is based on two nonnegative thresholds $k_{1,2}$ and $k_{2,1}$ and two positive queue-ratio parameters $r_{1,2}$ and $r_{2,1}$. We define two (weighted) queue-difference stochastic processes $D_{1,2}(t) \equiv Q_1(t) - r_{1,2}Q_2(t)$ and $D_{2,1}(t) \equiv r_{2,1}Q_2(t) - Q_1(t)$. As long as $D_{1,2}(t) \leqslant k_{1,2}$ and $D_{2,1}(t) \leqslant k_{2,1}$, agents may only serve customers from their designated class. (Ordinary FQR without thresholds corresponds to $r_{2,1} = r_{1,2}$ and $k_{1,2} = k_{2,1} = 0$.)

However, pool-2 agents are allowed to start serving class-1 customers when $D_{1,2}(t) > k_{1,2}$, provided that no pool-1 agents are still serving a class-2 customer. (We restrict attention to sharing in only one direction at a time, but either direction is possible.) Pool 2 is allowed to begin service as soon as no pool-1 agents are serving class-2 customers and $D_{1,2}(t) > k_{1,2}$. As soon as the first pool-2 agent is assigned to serve a class-1 customer, we drop the threshold $k_{1,2}$, but keep the other threshold $k_{2,1}$. Thus, once one-way sharing has been activated with pool 2 helping class 1, we use ordinary FQR with ratio parameter $r_{1,2}$: Upon service completion, a newly available type-2 agent serves the class-1 customer who has waited the longest if $D_{1,2}(t) > 0$; otherwise, the agent serves a customer from his own class. (There also is the other threshold $k_{2,1}$, but it will usually not be crossed during the overload incident.) Only one-way sharing in this direction will be allowed until either the class-1 queue becomes empty or the other difference process crosses the other threshold, i.e., when $D_{2,1}(t) > k_{2,1}$. As soon as either of these events occurs, newly available pool-2 agents are only assigned to class 2 and the threshold $k_{1,2}$ is reinstated and, similarly, in the other direction.

Even though we intend to drop the threshold $k_{1,2}$ when sharing is activated with pool 2 helping class 1 (in the manner just described), we consider a centering constant $\kappa_{1,2}$ after sharing, which can be interpreted as a threshold. Perry and Whitt (2009) show that in some cases it is actually optimal to use the *shifted FQR-T control*, i.e., keeping the queues at a fixed ratio centered about a constant. Such is the case, for example, when the holding cost is separable and quadratic, i.e., of the form $C(Q_1, Q_2) = C_1(Q_1) + C_2(Q_2)$, where $C_i(Q_i) = a_i + b_iQ_i + c_iQ_i^2$; this is proved in §EC.4 in Perry and Whitt (2009). In these cases the optimal relation between the queues is $Q_1 + r_{1,2}Q_2 = \kappa_{1,2}$ or $Q_1 + r_{2,1}Q_2 = \kappa_{2,1}$ for some $\kappa_{1,2}, \kappa_{2,1} \in \mathbb{R}$, depending on the direction of sharing; explicit formulas for the optimal ratios and centering constants appear in EC.11 and EC.12 of Perry and Whitt (2009). If $b_i = 0$ for $i = 1,2$, then the two centering constants take the form $\kappa_{1,2} = \kappa_{2,1} = 0$, and we have ordinary FQR once sharing has been activated in some direction.

More on the MSHT limiting regime. The MSHT regime is specified by considering a sequence of models indexed by n, here denoted by a superscript. The main idea is that the system scale should grow with n. Accordingly, we assume that the arrival rates and number of servers grow proportionally to n:

$$\frac{\lambda_i^{(n)}}{n} \to \bar{\lambda}_i \quad \text{and} \quad \frac{m_j^{(n)}}{n} \to \bar{m}_j \quad \text{as } n \to \infty, \tag{1}$$

where $\bar{\lambda}_i$ and $\bar{m}_j$ are positive constants for $i = 1,2$ and $j = 1,2$. The individual abandonment rates θ_i and service rates $\mu_{i,j}$ remain constant for all n. We add superscript (n) to all processes along the sequence of systems we consider, e.g., $Q_i^{(n)}(t)$ denotes the number of class-i customers in queue at time t in system n (having arrival rates $\lambda_1^{(n)}$ and $\lambda_2^{(n)}$ and $m_1^{(n)}$ and $m_2^{(n)}$ agents in the pools).

Because our model is overloaded, we will be considering a special case of the *efficiency-driven* (ED) MSHT regime; see Garnett et al. (2002) and Whitt (2004). For a Markovian I model, having one service pool, one customer class and customer abandonment, i.e., the $M/M/m+M$ model, we would be assuming that $\rho^{(n)} = \rho > 1$ for all n, where $\rho^{(n)} \equiv \lambda^{(n)}/n\mu$ is the traffic intensity in model n. With customer abandonment, the ED regime is quite practical because the queue lengths have proper steady-state distributions whenever the abandonment rates are positive. In this setting, we consider the scaled processes

$$\bar{Q}_i^{(n)}(t) \equiv \frac{Q_i^{(n)}(t)}{n} \quad \text{and} \quad \bar{Z}_{i,j}^{(n)}(t) \equiv \frac{Z_{i,j}^{(n)}(t)}{n}, \quad t \geqslant 0. \tag{2}$$

These scaled processes converge as $n \to \infty$, with

$$(\bar{Q}_i^{(n)}(t), \bar{Z}_{i,j}^{(n)}(t),\ i=1,2; j=1,2) \Rightarrow (\bar{Q}_i(t), \bar{Z}_{i,j}(t),\ i=1,2; j=1,2) \quad \text{as } n \to \infty, \tag{3}$$

where $\Rightarrow$ denotes convergence in distribution and the limit $(\bar{Q}_i(t), \bar{Z}_{i,j}(t),\ i=1,2; j=1,2)$ evolves as a deterministic ODE. The limit in (3) is referred to as a *functional weak* *law of large numbers* (FWLLN); it is established in Perry and Whitt (2011b).

Perry and Whitt (2011c) show that there is also an associated *functional central limit theorem* (FCLT) establishing associated stochastic limits, which serve as refinements of the fluid limits above. For these, we introduce the new scaled processes

$$\hat{Q}_i^{(n)}(t) \equiv \frac{Q_i^{(n)}(t) - n\bar{Q}_i(t)}{\sqrt{n}} \quad \text{and}$$

$$\hat{Z}_{i,j}^{(n)}(t) \equiv \frac{Z_{i,j}^{(n)}(t) - n\bar{Z}_{i,j}(t)}{\sqrt{n}}, \quad t \geqslant 0. \tag{4}$$

These scaled processes also converge as $n \to \infty$, with

$$(\hat{Q}_i^{(n)}(t), \hat{Z}_{i,j}^{(n)}(t),\ i = 1,2;\ j = 1,2) \Rightarrow (\hat{Q}_i(t), \hat{Z}_{i,j}(t),\ i = 1,2;\ j = 1,2) \quad \text{as } n \to \infty, \tag{5}$$

where the limit $(\hat{Q}_i(t), \hat{Z}_{i,j}(t),\ i = 1,2;\ j = 1,2)$ evolves as a stochastic (not deterministic) process.

3. The ODE Based on the AP

Overload scenarios. In this section we develop the fluid approximation, working with a single X model (not considering the MSHT regime). We do not know in advance, which class will experience the overload and need help from the other service pool. Indeed, the direction of sharing may switch in successive overload incidents. However, without loss of generality, when we consider the behavior of the system in one particular overload incident, under an unbalanced overload, we assume that class 1 is overloaded, and more so than class 2 if class 2 is also overloaded. Hence, we need only consider the queue-difference process $D_{1,2}(t)$, now denoted by $D(t) \equiv Q_1(t) - rQ_2(t)$.

When we say that class 1 is overloaded, we mean that $\lambda_1 > m_1\mu_{1,1}$. There are two cases for the less-loaded class 2 after sharing: We may either have class 2 also overloaded, but less so than class 1, or class 2 underloaded. We will primarily be focusing on the fully overloaded case, in which class 2 is overloaded after sharing. That can occur in two ways. First, class 2 might be overloaded by itself, before helping class 1. That occurs if $\rho_2 \equiv \lambda_2/m_2\mu_{2,2} > 1$. Alternatively, class 2 may be underloaded before sharing, but become overloaded on account of the sharing.

Approximation when fully overloaded. We now develop the approximation for the transient behavior of the CTMC $\{(Q_i(t), Z_{i,j}(t);\ i = 1,2;\ j = 1,2)\colon t \geqslant 0\}$ during an overload incident in the fully overloaded case. We start when the overload begins, at the instant the arrival rates change. The ODEs should apply to all possible initial conditions, but the standard case is for the system to be initially in steady state with the two service pools operating independently at normal levels. The sudden shift in the arrival rates causes the system to go through two transient periods. In the first transient period, the two systems continue to operate independently, with each responding to its own new arrival rate. In the fully overloaded case being considered, the first transient period ends and the second transient period begins after $D(t)$ exceeds its threshold and sharing is initiated, with all servers in both pools busy. That is when the AP begins to operate. The system evolves in this second transient period approaching the steady-state associated with the overload. In this section we are focusing on the second transient period; see Appendix B of Perry and Whitt (2011a) for discussion of the first transient period.

The averaging principle (AP). We can exploit SSC to deduce the relation $Q_1(t) = rQ_2(t) + \kappa$ for each t in the fully overloaded case. However, it is evident that SSC does not actually occur in such a simple way. Instead, the queue-difference process $D(t) \equiv Q_1(t) - rQ_2(t)$ oscillates around the centering constant κ. The key observation is that the queue-difference process $D(t)$ moves back and forth across the boundary κ relatively quickly, because it has a strong drift pointing toward κ on both sides (under typical overload conditions). These boundary crossings occur in a faster time scale than the relative changes in the other processes under consideration. Even though all processes move due to arrivals and service completions (which are happening quickly because the system is large), the relative changes of the processes $Q_i(t)$ and $Z_{i,j}(t)$ over short time intervals are small due to their size, which is of the same order as the number of servers. (For that reason, a continuous fluid approximation for these processes is appropriate.) In contrast, $D(t)$ *does not grow with the system's size*, and stays close to the boundary κ throughout. Hence, over very short time intervals, $D(t)$ moves rapidly between the two regions $(-\infty, \kappa]$ and (κ, ∞), with its speed growing proportionally to the size of the system. From the asymptotic perspective, $Q_i(t)$ and $Z_{i,j}(t)$ evolve with an $O(1)$ clock, whereas $D(t)$ evolves with an $O(1/n)$ clock when the arrival rate is of order n.

In particular, we conclude that $D(t)$ approximately reaches a time-dependent steady state instantaneously at each time t, where that steady-state distribution depends on the time-dependent quantities $Q_i(t)$ and $Z_{i,j}(t)$ (t fixed); i.e., there is an AP. For each $t \geqslant 0$, let $D_t(\infty)$ denote a random variable with that time-dependent steady-state distribution. We will then exploit the time-dependent probabilities $\pi_{1,2}(X(t)) \equiv P(D_t(\infty) > \kappa)$. To obtain the probability distribution of the steady-state random variable $D_t(\infty)$, we introduce a new stochastic process, the *fast-time-scale process* (FTSP) $D_t \equiv \{D_t(s)\colon s \geqslant 0\}$, which is the process $\{D(t+s)\colon s \geqslant 0\}$, initialized at $D(t)$, but with the transition rates of the stochastic process D under the extra condition that $(Q_1, Q_2, Z_{1,2})$ remain fixed at their values at time t.

Based on the AP, the FTSP D_t is a pure-jump continuous-time Markov process (CTMP), with state space $\{k + rj\colon k \in \mathbb{Z},\ j \in \mathbb{Z}\}$. with transition rates that depend only on the fluid-model state at time t. There are four possible transitions in each state: ±1 and $\pm r$. We obtain simplification without practical sacrifice by assuming that r is rational. For rational $r \equiv j/k$, the FTSP is a CTMC on the state space $\{j/k\colon j \in \mathbb{Z}\}$. We multiply by k to make all the states integers. Moreover, then the CTMC can be represented as a homogeneous quasi-birth-and-death (QBD) process, as in Definition 1.3.1 and §6.4 of Latouche and Ramaswami (1999). For each t, we can apply the logarithmic reduction algorithm in §8.7 of Latouche and Ramaswami (1999) to efficiently calculate the steady-state distribution of D_t, i.e., the distribution of $D_t(\infty)$. As a consequence, we can calculate the desired probabilities $\pi_{1,2}(X(t))$ given any state vector $X(t) \equiv (Q_1(t), Q_2(t), Z_{1,2}(t))$.

We now specify the transition rates of the CTMC D_t given the time t and the state $X(t)$, using the integer state space. Let $\lambda_+^{(j)}(m, X(t))$, $\lambda_+^{(k)}(m, X(t))$, $\mu_+^{(j)}(m, X(t))$, and $\mu_+^{(k)}(m, X(t))$ be the transition rates of the FTSMC D_t for transitions of $+j$, $+k$, $-j$, and $-k$, respectively, when $D_t(s) = m > \kappa$. Similarly, we define the transitions when $D_t(s) = m \leqslant \kappa$: $\lambda_-^{(j)}(m, X(t))$, $\lambda_-^{(k)}(m, X(t))$, $\mu_-^{(j)}(m, X(t))$, and $\mu_-^{(k)}(m, X(t))$.

First, for $D_t(s) = m \in (-\infty, \kappa]$, the upward rates are

$$\lambda_-^{(k)}(m, X(t)) = \lambda_1, \quad \text{and}$$
$$\lambda_-^{(j)}(m, X(t)) = \mu_{1,2} Z_{1,2}(t) + \mu_{2,2} Z_{2,2}(t) + \theta_2 Q_2(t), \qquad (6)$$

corresponding, first, to a class-1 arrival and, second, to a departure from the class-2 queue, caused by a type-2 agent service completion (of either customer type) or by a class-2 customer abandonment. Similarly, the downward rates are

$$\mu_-^{(k)}(m, X(t)) = \mu_{1,1} Z_{1,1}(t) + \theta_1 Q_1(t) \quad \text{and}$$
$$\mu_-^{(j)}(m, X(t)) = \lambda_2, \qquad (7)$$

corresponding, first, to a departure from the class-1 customer queue, caused by a class-1 agent service completion or by a class-1 customer abandonment, and, second, to a class-2 arrival. Next, for $D_t(s) = m \in (\kappa, \infty)$, we have upward rates

$$\lambda_+^{(k)}(m, X(t)) = \lambda_1 \quad \text{and} \quad \lambda_+^{(j)}(m, X(t)) = \theta_2 Q_2(t), \qquad (8)$$

corresponding, first, to a class-1 arrival and, second, to a departure from the class-2 customer queue caused by a class-2 customer abandonment. The downward rates are

$$\begin{aligned} &\mu_+^{(k)}(m, X(t)) \\ &= \mu_{1,1} Z_{1,1}(t) + \mu_{1,2} Z_{1,2}(t) + \mu_{2,2} Z_{2,2}(t) + \theta_1 Q_1(t) \quad \text{and} \\ &\mu_+^{(j)}(m, X(t)) = \lambda_2, \end{aligned} \qquad (9)$$

corresponding, first, to a departure from the class-1 customer queue, caused by (i) a type-1 agent service completion, (ii) a type-2 agent service completion (of either customer type), or (iii) by a class-1 customer abandonment and, second, to a class-2 arrival.

We conclude the definition of the FTSP by noting that great simplification occurs in the special case $r = 1$, because then the CTMC reduces to a simple birth-death (BD) process instead of a QBD process. Then it is easy to calculate $\pi_{1,2}(X(t))$; see Theorem 6.2 of Perry and Whitt (2011a).

The ODE. The fluid approximation is a solution to an ODE and, in particular, it is a differentiable function. Its derivative at each time t approximates the instantaneous rates of $X(t)$, which is a CTMC, provided all agents are working and there are no class-2 customers in pool 1 (which is what we assume). We have just observed that the rate of change of the FTSP D_t depends on (i) the state $X(t)$ and (ii) whether or not $D_t(s) > \kappa$. In the same way, the rates of the CTMC X at time t depend on (i) $X(t)$ itself and (ii) the state of $D(t)$. Now, the deterministic fluid approximation of the evolution of $X(t)$, we let the rates (i.e., derivatives) depend on (i) $X(t)$ itself and (ii) the steady-state probability $\pi_{1,2}(X(t)) = P(D_t(\infty) > \kappa)$.

First, given $Z_{i,j}(t)$ and $\pi_{i,j}(X(t))$, we obtain ODE's for the two queue-length processes. Let $\dot{Q}_i \equiv \dot{Q}_i(t)$ denote the derivative of Q_i evaluated at t. The derivative $\dot{Q}_1(t)$ equals the rate of increase minus its rate of decrease. The rate of increase is simply the arrival rate to customer queue 1, λ_1. The rate of decrease is more complicated. First, there is the rate of abandonment from queue 1, which is $Q_1(t)\theta_1$. Second, there is the rate of decrease from queue 1 due to service completions by servers who will next take customers from queue 1, which depends on the state of the queue-difference stochastic process. Exploiting the AP, we will not focus on the actual state of the queue-difference process, but instead focus on the average state, assuming that the queue-difference process oscillates relatively rapidly compared to the other processes. We thus assume that a proportion $\pi_{1,2}(X(t))$ of the time that the queue-difference exceeds the shifting constant κ. That portion of the decrease rate is $\pi_{1,2}(X(t))(Z_{1,2}(t)\mu_{1,2} + Z_{2,2}(t)\mu_{2,2})$. There will be corresponding, but different, rates of decrease for the proportion of time $1 - \pi_{1,2}(X(t))$ that the queue difference is less than or equal to κ. That reasoning leads to the system of three ODEs

$$\begin{aligned} \dot{Q}_1(t) &\equiv \lambda_1 - m_1\mu_{1,1} - \pi_{1,2}(X(t)) \\ &\qquad \cdot [Z_{1,2}(t)\mu_{1,2} + Z_{2,2}(t)\mu_{2,2}] - \theta_1 Q_1(t) \\ \dot{Q}_2(t) &\equiv \lambda_2 - (1 - \pi_{1,2}(X(t))) \\ &\qquad \cdot [Z_{2,2}(t)\mu_{2,2} + Z_{1,2}(t)\mu_{1,2}] - \theta_2 Q_2(t) \\ \dot{Z}_{1,2}(t) &\equiv \pi_{1,2}(X(t)) Z_{2,2}(t)\mu_{2,2} \\ &\qquad - (1 - \pi_{1,2}(X(t))) Z_{1,2}(t)\mu_{1,2}. \end{aligned} \qquad (10)$$

More compactly, we have a single three-dimensional ODE with the general form $\dot{X}(t) = \Psi(X(t), t)$ for a function Ψ. In addition, our ODE is *autonomous* (or *time invariant*) because $\Psi(X(t), t) \equiv \Psi(X(t))$. An autonomous ODE does not depend explicitly on the time-argument t, and its behavior is invariant to shifts in the time origin. Thus, we propose the autonomous ODE

$$\begin{aligned}\dot{X}(t) &\equiv (\dot{Q}_1(t), \dot{Q}_2(t), \dot{Z}_{1,2}(t)) = \Psi(X(t))\\ &\equiv \Psi(Q_1(t), Q_2(t), Z_{1,2}(t)), \quad t \geqslant 0, \end{aligned} \tag{11}$$

where Ψ: $[0,\infty)^2 \times [0, m_2] \to \mathbb{R}^3$ is displayed via (10) above. The derivatives in (10) are evident given the transition rates of the CTMC, given that we replace the CTMC by an ODE and invoke the AP.

A more systematic derivation of the ODE (10), involving the asymptotic approach, appears in Perry and Whitt (2011a, b). In particular, Theorem 5.2 in Perry and Whitt (2011a) proves that there exists a unique solution to the ODE that is continuous and differentiable almost everywhere. The setting considered in Perry and Whitt (2011a) is much more general than here, and the unique solution is shown to exist in the full three-dimensional state space (not only in the two-dimensional state space where the AP operates and SSC of the queues occurs). Thus, there exists a unique solution to the ODE for any set of parameters that puts the system into overload (see Assumption A in Perry and Whitt 2011a). Building on that existence and uniqueness result, Theorem 6.1 in Perry and Whitt (2011b) proves that the solution to that ODE is achieved as the MSHT fluid limit for the sequence of stochastic systems.

What is important for us here is that we can apply standard iterative algorithms for solving ODEs to solve (11), where we calculate $\pi_{1,2}(X(t))$ at each step. We used the classical forward Euler algorithm for the ODE together with the logarithmic reduction algorithm for QBDs from Latouche and Ramaswami (1999); additional details are provided in §§6 and 11 of Perry and Whitt (2011a).

4. Validating the Transient Approximation Through Simulation Experiments

We now provide evidence that our proposed approximation is effective for the transient behavior. Accordingly, in this section we compare numerical results for the transient behavior of the fluid model, based on our algorithm from Perry and Whitt (2011a), to simulation estimates of the actual performance measures in the original queueing model. This will show that the transient approximations are computable and sufficiently accurate for engineering applications. We also show that the deterministic fluid model does not capture important stochastic fluctuations unless the scale is very large, but the fluid model provides remarkably accurate approximations for the mean values of the key queueing processes, $Q_1(t)$, $Q_2(t)$, and $Z_{1,2}(t)$, provided that the scale is not too small.

In order to demonstrate the MSHT limits in the ED regime described in §2, we report results for scaled processes, as in (2), for several values of n. We will then be confirming the FWLLN in (3) via the simulation. Our simulation examples throughout the paper will have parameters related to a *base case* that we consider here as well. It has several parameters depending on n: $m_i \equiv m_i^{(n)} = n$, $\lambda_1 \equiv \lambda_1^{(n)} = 1.3n$, $\lambda_2 \equiv \lambda_2^{(n)} = 0.9n$, and $\kappa \equiv \kappa^{(n)}$. Here we take $\kappa^n = 0$, but we will later also consider a positive κ, specifically, $\kappa \equiv \kappa^n = 0.1n$. The other model parameters are independent of n: $\theta_1 = \theta_2 = 0.2$, $\mu_{1,1} = \mu_{2,2} = 1.0$, and $\mu_{1,2} = \mu_{2,1} = 0.8$. The arrival rates are chosen to put class 1 in a focused overload, whereas class 2 is initially normally loaded or slightly underloaded, but becomes overloaded too after the sharing. The rest of the parameters are chosen to make a symmetric model, where serving the other class is less efficient. We use the FQR-T control with ratio parameter $r = 0.8$; this makes the QBD matrices be as in (6.5) and (6.6) of Perry and Whitt (2011a), following the general structure in §§6.1 and 6.2 there; the algorithm is given in §11 there.

We have in mind large-scale applications, e.g., with $n \geqslant 50$, but to test the limits of the approximations, we also consider smaller systems. Specifically, we consider the three cases: $n = 10$, $n = 25$, and $n = 100$, initialized empty. Because the processes are scaled, they all have the same fluid approximation. For each n, we ran 1,000 independent replications, sampling each of the 1,000 simulated sample paths every $h \equiv 0.01$ time units over the time interval $[0, T] = [0, 50]$. This gives 5,001 sample points for each replication.

Figures 2–4 show the fluid approximation together with simulation estimates of the time-dependent mean values for each n, specifically, the averages of the 1,000 observed values of three scaled processes $\bar{Q}_i^{(n)}(t) \equiv n^{-1} Q_i^{(n)}(t)$, $i = 1, 2$, and $\bar{Z}_{1,2}^{(n)}(t) \equiv n^{-1} Z_{1,2}^{(n)}(t)$ at each of the 5,001 sample points. Figure 5 shows one sample path of $\{\bar{Q}_1^{(n)}(t)\colon 0 \leqslant t \leqslant 50\}$, when $n = 100$, together with the fluid approximation, to show the typical stochastic fluctuations. These

Figure 2. A comparison of simulation estimates of $E[\bar{Q}_1^{(n)}(t)]$ for $n = 10, 25, 100$ to the fluid approximation in the base case.

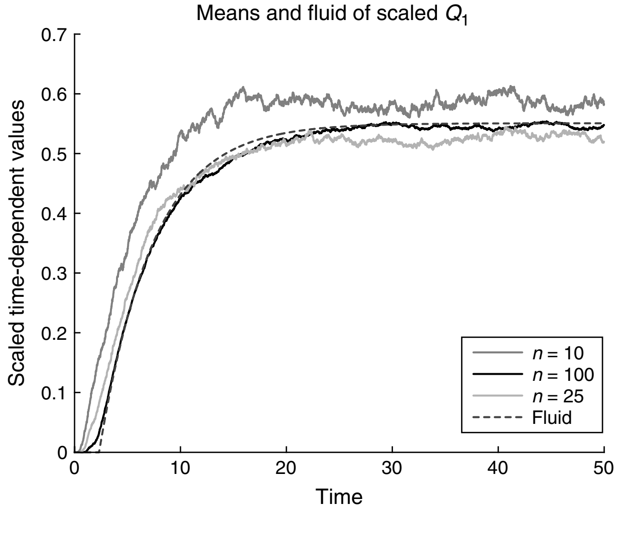


Figure 3. A comparison of simulation estimates of $E[\bar{Q}_2^{(n)}(t)]$ for $n = 10, 25, 100$ to the fluid approximation in the base case.

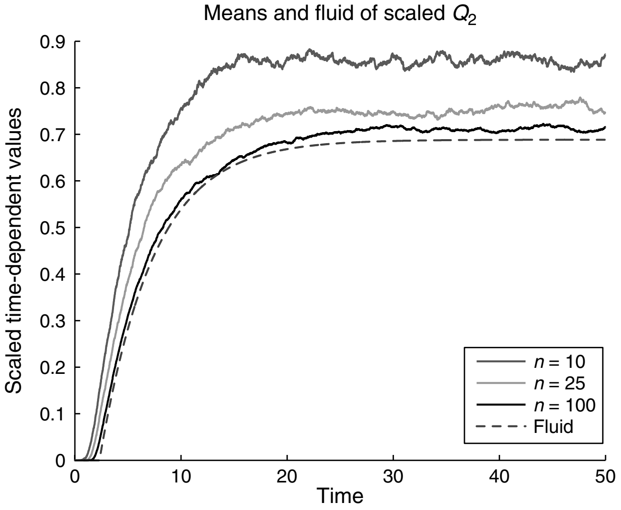
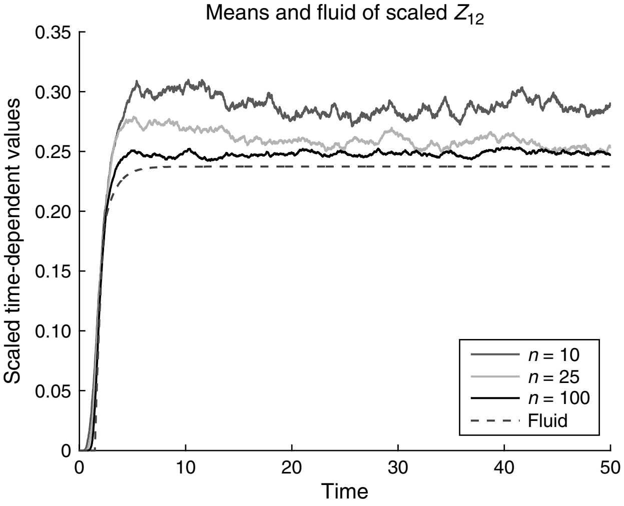


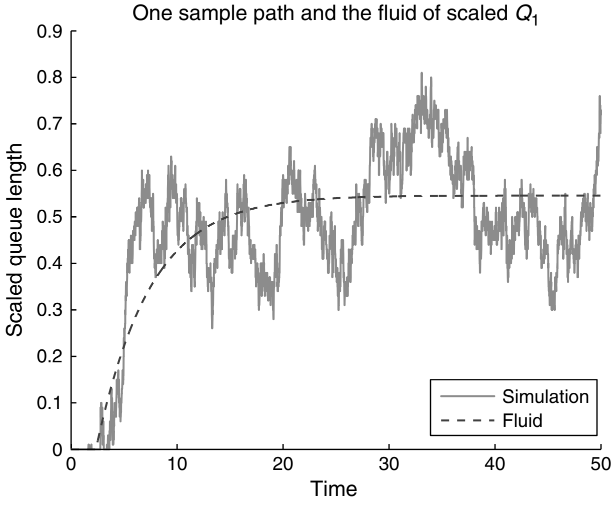
Figure 4. A comparison of simulation estimates of $E[\bar{Z}_{1,2}^{(n)}(t)]$ for $n = 10, 25, 100$ to the fluid approximation in the base case.



stochastic fluctuations are the reason for using a large number of replications in order to accurately estimate the mean values at each point along the sample path. The statistical precision of the estimators is directly visible in the plots, because the processes are effectively in steady state in the second half of the time interval $[0, 50]$. As n grows larger, the impact of these fluctuations decreases; they are of order $1/\sqrt{n}$ by (5). The stochastic fluctuations show the importance of the diffusion refinements in §7.

Consistent with the FWLLN in (3), the larger the system, the better the fluid approximates the means. The figures clearly show that $n \geqslant 100$ is "large enough" in the sense that the simulated means are extremely close to the fluid approximation. Even a relatively small system, with only 25 agents in each pool, is approximated quite well by the fluid. However, the fluid approximation is quite rough when $n = 10$. There is approximately 25% difference between the fluid and the means of $\bar{Q}_2^{(n)}(t)$ when $n = 10$.

Figure 5. A comparison of one sample path of $\bar{Q}_1^{(n)}(t)$ when $n = 100$ to the fluid approximation in the base case.



Nevertheless, the fluid approximation is useful even for small systems, because the shape of the curves of the simulation means for $n = 10$ is the same as the shape of the fluid curve; in particular, the rate of convergence to steady state is about the same in all systems. Because the fluid approximation was shown to converge exponentially fast to steady state in §9 of Perry and Whitt (2011a), we see that the same must be true, approximately, for the queueing system even for a quite small numbers of servers.

5. Stochastic Refinements to the Steady-State Fluid Approximation

In this section we present two stochastic refinements to the deterministic fluid-model approximations for the steady-state quantities Q_i and $Z_{1,2}$ describing performance during the overload, assuming shifted FQR is used. The first exploits the AP to determine the average queue difference for the fully overloaded case. The second develops a birth-and-death-process (BD) approximation for the steady-state queue length Q_1 in the spare capacity case.

For reference, we refer to the steady-state approximation based on the simple flow balance from Perry and Whitt (2009). Based on the argument there, we can find the three variables Q_1, Q_2, and $Z_{1,2}$ by solving the following two equations in two unknowns (Q_1 and $Z_{1,2}$):

$$Q_1 = \frac{\lambda_1 - (m_1\mu_{1,1} + Z_{1,2}\mu_{1,2})}{\theta_1} \quad \text{and}$$

$$Q_2 = \frac{Q_1 - \kappa}{r} = \frac{\lambda_2 - (m_2 - Z_{1,2})\mu_{2,2}}{\theta_2}. \tag{12}$$

This simple approximation can also be derived from the ODE. We can directly apply the ODE for $Z_{1,2}(t)$ to find $\pi_{1,2}$ by noting that in steady state $\dot{Z}_{1,2}(t) = 0$. Thus,

$$\pi_{1,2} = \frac{Z_{1,2}\mu_{1,2}}{Z_{1,2}\mu_{1,2} + (m_2 - Z_{1,2})\mu_{2,2}}. \tag{13}$$

Setting $\dot{X}(t) = 0$ in (11) and applying (13) yields (12) above.

5.1. The Average Difference E[D] in the Fully Overloaded Case

We have observed that SSC does not happen exactly; we do not get precisely $Q_1 = rQ_2 + \kappa$. Instead, the queue-difference process $D(t)$ oscillates around the centering constant κ. We can apply the AP to find an approximating steady-state distribution of $D(t)$ by treating it as an FTSP. Let D denote a random variable with the limit of these steady-state distributions as $t \to \infty$.

We propose refining our fluid approximation for the steady-state distribution by replacing the target difference κ in (12) by the mean $E[D]$. To find $E[D]$, we solve the balance equations of the FTSP and then take the mean

$$E[D] = \sum_{j=-\infty}^{\infty} jP(D = j). \tag{14}$$

Because the drifts tend to point strongly toward the centering constant κ, it usually suffices to perform the sum for $\kappa - 20 \leqslant j \leqslant \kappa + 20$.

We now obtain our refined approximation, assuming that the queue difference is $E[D]$ instead of κ. The calculation of $E[D]$ can be easily done if Q_1, Q_2, and $Z_{1,2}$ are known. Because they depend on the value $E[D]$, we need to solve for them simultaneously. To do that, we propose an iterative algorithm that solves the *three equations*

$$\begin{aligned} Q_1 &= \frac{\lambda_1 - (m_1\mu_{1,1} + Z_{1,2}\mu_{1,2})}{\theta_1}, \\ Q_2 &= \frac{Q_1 - E[D]}{r} = \frac{\lambda_2 - (m_2 - Z_{1,2})\mu_{2,2}}{\theta_2}, \\ E[D] &= \sum_{j=-k_{2,1}}^{\infty} jP(D = j). \end{aligned} \tag{15}$$

For the iterative procedure, it is natural to start with the values of Q_1, Q_2, and $Z_{1,2}$ obtained from (12), and then calculate the distribution of D and $E[D]$. We can then obtain new values of Q_1, Q_2, and $Z_{1,2}$ by solving (12) again with $E[D]$ replacing κ. We then can keep iterating. Experience indicates that this iteration consistently converges in a few iterations (typically only two), yielding the solution to (15).

5.2. A BD-Process Refinement for the Spare-Capacity Case

For the case in which queue 2 has spare capacity, we now develop another refinement, obtaining a nondegenerate approximation for the distribution of Q_1. In this case, because of the available agents in pool 2, as soon as Q_1 exceeds the centering constant κ, an idle pool-2 agent serves a customer from class 1. Thus, it is evident that we must have $Q_1 \leqslant \kappa$. Of course, the fluid approximation is just $Q_1 \approx \kappa$.

Because of the averaging principle, it is not hard to estimate the approximate distribution of Q_1. To do so, we observe that we can regard the class-1 queue as evolving below the level $\kappa_{1,2}$ by itself as a BD process. When the queue length is j, the birth rate is a constant λ_1, whereas the death rate is approximately $m_1\mu_{1,1} + \theta_1 j$. (Queue 2 plays no role.) For the reason given, the birth rate is 0 when the queue is at κ. The death rate should be small when the queue length is small. For the approximation to be good, we do not want Q_1 to spend much time at very low levels, like 1 or 0. That can be verified approximately by looking at the approximate BD steady-state distribution. In any case, we let the death rate be 0 when the queue length is 0. Our refined approximation for the distribution of Q_1 is the steady-state distribution of this finite-state BD process.

Because $Q_1^{\text{alone}} = (\lambda_1 - m_1\mu_{1,1})/\theta_1 > \kappa$, the birth rate always exceeds the death rate here. Indeed, the BD process here for $\kappa - Q_1(t)$ is stochastically bounded above by the queue-length process in an $M/M/1/\kappa$ queue, where κ serves as the size of a finite waiting room. If we take the asymptotic perspective in §2, this stochastic bound shows that the difference $\kappa - Q_1$ should be of order $O(1)$ as $n \to \infty$. Hence, this adjustment should be asymptotically negligible in both the diffusion scale ($\sqrt{n}$) and the fluid scale (n). However, the refinement can help in actual examples, even large ones with 1,000 servers in each pool.

As a refined deterministic fluid approximation, we use the mean value of the steady-state distribution of the BD process here. However, by this method, we also obtain an estimate for the variance and the entire distribution of Q_1. The observed $M/M/1$ structure indicates that the distribution of $\kappa - Q_1(t)$ should be approximately a truncated geometric distribution. That is quite different from the approximate normal distribution we derive for the fully overloaded case in §7.

6. Simulation Experiments to Evaluate the Steady-State Mean Values

The fully overloaded case. We have developed deterministic fluid approximations for the steady-state mean values in the fully overloaded case via the solutions to the two equations in (12) and the three equations in (15). We now compare these approximations to simulation estimates. In order to use the simulation to substantiate the conjectured

Table 1. A comparison of the basic fluid approximations based on two equations in (12) and its refinement based on the three equations in (15) with simulation results in the base case, having $m_1 = m_2 = 1.0n$, $\lambda_1 = 1.3n$, $\lambda_2 = 0.9n$, $\mu_{1,1} = \mu_{2,2} = 1.0$, $\mu_{1,2} = \mu_{2,1} = 0.8$, $\theta_1 = \theta_2 = 0.2$, and $\kappa = 0.1n$ (rounding up to the nearest integer if necessary).

	$n = 25$			$n = 100$			$n = 400$		
Perf. meas.	2 Eq.	3 Eq.	Sim.	2 Eq.	3 Eq.	Sim.	2 Eq.	3 Eq.	Sim
$E[Q_1]$	16.6	14.4	15.7 ± 0.3	65.6	63.1	63.6 ± 1.9	262.2	259.7	258.3 ± 5.0
$E[Q_1/n]$	0.656	0.575	0.629 ± 0.013	0.656	0.631	0.636 ± 0.019	0.656	0.649	0.646 ± 0.013
$E[Q_2]$	13.6	16.4	15.9 ± 0.4	55.6	58.6	58.6 ± 1.8	222.2	225.3	223.9 ± 5.0
$E[Q_2/n]$	0.556	0.656	0.636 ± 0.016	0.556	0.586	0.586 ± 0.018	0.556	0.563	0.560 ± 0.013
$E[D]$	—	-2.0	-0.2 ± 0.3	—	4.6	5.0 ± 0.1	—	34.4	34.4 ± 0.04
$\kappa - E[D]$	—	5.0	3.2 ± 0.3	—	5.4	5.0 ± 0.1	—	5.6	5.6 ± 0.04
$E[Z_{1,2}]$	5.3	5.8	5.6 ± 0.1	21.1	21.7	21.9 ± 0.04	84.4	85.1	84.2 ± 1.2
$E[Z_{1,2}/n]$	0.211	0.231	0.224 ± 0.003	0.211	0.217	0.219 ± 0.004	0.211	0.213	0.210 ± 0.003

stochastic process limits in §2, we choose parameters corresponding to scaled systems, indexed by n, letting n take the values 25, 100, and 400. We have considered much larger n, such as $n = 1{,}000$, but from the results for $n = 400$, we see that accurate results will be obtained for all n larger than 400.

We consider the base case, introduced in §4, with $r = 1$. This makes the model symmetric and reduces the fast-scale MP to a BD process. In the online version we present corresponding results for asymmetric models. In all our simulation experiments, we used five independent runs, each with 300,000 arrivals. We report averages together with the half-widths of the 95% confidence intervals, based on a t statistic with four degrees of freedom. Simulation results for the base case above are presented in Table 1 below. Table 1 shows both the steady-state mean values and the associated scaled values (i.e., divided by n). The unscaled values helps us evaluate the performance of the actual system, whereas the scaled values show the convergence of the stochastic process limits in (3). Table 1 clearly shows that the level of accuracy grows as n gets larger, but even for relatively small systems, the fluid approximation gives reasonable results.

Table 1 also gives the approximation for the steady-state mean of the unscaled weighted-difference process $D(t)$ as developed in §5.1, and compares it to simulation results. The sixth row in the table is especially insightful. It shows that $E[D]$ is about the same distance from $\kappa_{1,2}$ for each n, thus strengthening our claim that $D(t)$ should have fluctuations of order $O(1)$ as $n \to \infty$. In closing, we remark that we rounded up the centering constant κ to the nearest integer when $n = 25$; i.e., we used $\kappa = 3$ when $n = 25$. In the table we show the fluid solution using $\kappa = 2.5$ so as to make the scaled fluid solutions uniform. However, the solution using $\kappa = 3$ is similar.

Independent cases. One of our objectives is to avoid sharing without unbalanced overloads. That occurs in two scenarios: (i) under normal loads and (ii) under balanced overloads. In both of these cases, our FQR-T control makes the X model operate approximately as two independent $M/M/n + M$ systems, each operating in the QD or QED regime in the first scenario (depending on the actual load of each queue), or the ED regime in the second scenario. We present supporting simulation results in the online version.

The spare-capacity case. For the spare capacity case, we modify the base case above to make queue-1 overloaded, whereas pool-2 has enough spare capacity to potentially serve all the extra class-1 customers. As before, we just change the arrival rates—in this case, to $\lambda_1 = 1.1n$ and $\lambda_2 = 0.8n$.

It is easy to see that pool 2 has spare capacity (in the fluid scale). We can analyze the available capacity from this deterministic fluid-approximation perspective as follows: First, we observe that class 1 has an extra arrival rate of $0.1n$, whereas pool 2 has $0.2n$ "extra" service rate, assuming that $0.8n$ servers are enough to take care of all the class-2 arrivals. Because pool-2 agents serve class-1 customers at rate $\mu_{1,2} = 0.8$, we initially estimate that we need to have at least $0.125n$ pool-2 agents working with class-1 customers. However, upon further analysis, we see that the number of pool-1 agents needed is actually less than that, because queue 1 will stabilize at the centering constant $\kappa = 0.1n$, and thus $\theta_1 Q_1 = 0.02n$ class-1 customers will abandon. Hence, only about $0.105n$ pool-2 agents should be needed to serve class 1. In any case, pool 2 has spare capacity.

We compare the approximation from §5.2 with simulation results in Table 2. Our initial approximation for Q_1 is κ, but that is not shown in Table 2. Instead, we only show the BD refinement from §5.2. (The cruder approximation would yield values of 2.5, 10.0, and 40.0 in the first row.) We see that the refined approximation is much better for large n. For the approximation of $Z_{1,2}$, we use

$$Z_{1,2} = \frac{\lambda_1 - m_1 \mu_{1,1} - \kappa\theta_1}{\mu_{1,2}}. \tag{16}$$

We obtain (16) using the flow balance reasoning of Perry and Whitt (2009) by observing that we achieve that value κ for Q_1 if and only if $Z_{1,2}$ serves to balance the rate in

Table 2. A comparison of the approximation for the steady-state performance measures in the spare-capacity case with simulation results.

	$n=25$		$n=100$		$n=400$	
Perf. meas.	Approx.	Sim.	Approx.	Sim.	Approx.	Sim.
$E[Q_1]$	1.1	3.3 ± 0.1	5.2	6.4 ± 0.6	29.0	30.1 ± 0.5
$E[Q_1/n]$	0.04	0.13 ± 0.00	0.05	0.06 ± 0.01	0.07	0.07 ± 0.00
$E[Q_2]$	0	3.4 ± 0.05	0	2.7 ± 0.5	0	1.0 ± 0.2
$E[Q_2/n]$	0	0.14 ± 0.00	0	0.027 ± 0.005	0	0.003 ± 0.000
$E[Z_{1,2}]$	2.5	3.9 ± 0.1	10.0	12.2 ± 0.5	40.0	43.4 ± 1.2
$E[Z_{1,2}/n]$	0.100	0.156 ± 0.007	0.100	0.122 ± 0.007	0.100	0.108 ± 0.003

Note. The arrival rates are now $\lambda_1 = 1.1n$ and $\lambda_2 = 0.8n$.

and rate out at queue 1. Because the rate into queue 1 is λ_1, whereas the rate out is $m_1\mu_{1,1} + \kappa\theta_1 + Z_{1,2}\mu_{1,2}$, we obtain (16). In order for queue 2 to be empty with the rate into queue 2 being λ_2, which is less than or equal to the maximum rate out of queue 2, which is $\mu_{2,2}(m_2 - Z_{1,2})$, to have $Q_2 = 0$ along with $Q_1 = \kappa$, we necessarily have $Z_{1,2} < m_2$.

7. A Diffusion Process Refinement

In the fully overloaded case, we now go beyond the deterministic fluid approximation to obtain a diffusion process refinement, which yields a nondegenerate approximation for the steady-state distribution of the two queue lengths. The approximating distribution is bivariate normal, where the means are the previous fluid approximations. In addition, the approximating correlation is 1 and the variances are

$$\mathrm{Var}(Q_1) \approx \frac{r^2(\lambda_1+\lambda_2)}{(1+r)(r\theta_1+\theta_2)} \quad \text{and}$$
$$\mathrm{Var}(Q_2) \approx \frac{(\lambda_1+\lambda_2)}{(1+r)(r\theta_1+\theta_2)}. \tag{17}$$

A special case. We base our approximation on a special case for which we can easily do the asymptotic analysis exactly, and then we extend the approximation heuristically to other cases. The special case has $\theta_1 = \theta_2$ and $\mu_{1,2} = \mu_{2,2}$ (with class 1 overloaded as usual). Under those additional assumptions, the total queue length $Q_s(t) \equiv Q_1(t) + Q_2(t)$ behaves the same as the queue length in the $M/M/m+M$ model in the ED regime, as analyzed in Whitt (2004). In this special case, we can directly obtain a FCLT like (5) for the total queue-length stochastic process, centered about the steady-state fluid limit. From Whitt (2004), we see that the limit is an Ornstein-Uhlenbeck diffusion process with infinitesimal mean $m(x) = -\theta_1 x$ and infinitesimal variance $\sigma^2 \equiv \sigma^2(x) = 2(\lambda_1 + \lambda_2)$. That diffusion process has a normal steady-state distribution. We invoke SSC to treat the individual queue lengths; that yields the correlation 1.

Here are additional details: Because the system is fully overloaded, as an approximation we assume that all the agents are busy all the time. (That is asymptotically correct in the MSHT limit.) Thus, the departure rate by service completion has the constant value $m_1\mu_{1,1} + m_2\mu_{2,2}$. The assumption that $\mu_{1,2} = \mu_{2,2}$ implies that it does not matter which class the type-2 agents are serving. Because the total arrival process is a superposition of two independent Poisson processes, the total arrival process is directly a Poisson process with rate $\lambda_1 + \lambda_2$. Finally, because $\theta_1 = \theta_2$, there is a common abandonment rate for both classes.

A heuristic refinement. Now we heuristically extend this same tractable OU approximation with a normal steady-state distribution to more general cases. First, when $\mu_{1,2} \neq \mu_{2,2}$, we again act as if all agents are busy all the time. The total service rate at time t is then $m_1\mu_{1,1} + Z_{1,2}(t)\mu_{1,2} + (m_2 - Z_{1,2}(t))\mu_{2,2}$. To obtain the desired constant rate, we act as if $Z_{1,2}(t)$ is constant, assuming its deterministic steady-state fluid approximation. This is a heuristic approximation, because we are ignoring the stochastic fluctuations in $Z_{1,2}$. Experiments show that this simple approximation works pretty well, but as $n \to \infty$ in the ED regime the infinitesimal mean of the scaled queue-length process does in fact depend on the stochastic behavior of the scaled version of the stochastic process $Z_{1,2}$ (as we would expect); i.e., simulations show that this heuristic extension is *not* asymptotically correct as $n \to \infty$, but it is a useful approximation.

We also treat the abandonments in a similar way when $\theta_1 \neq \theta_2$. We will approximate by a constant abandonment rate applying to all customers. For this step we also will invoke SSC (ignoring the difference) and assume that $Q_1(t) \approx rQ_s(t)/(1+r)$ (and similarly for Q_2). Thus, our approximating constant abandonment rate to apply to the total queue length is $\theta \approx (r\theta_1/(1+r)) + (\theta_2/(1+r))$. With the new approximating total service rate and average abandonment rate, we again are in the domain of an OU approximation, with normal steady-state distribution. Paralleling our previous analysis, we obtain a new approximate variance for the total queue length,

$$\mathrm{Var}(Q_s) \approx \frac{(1+r)(\lambda_1+\lambda_2)}{(r\theta_1+\theta_2)}. \tag{18}$$

Then SSC again gives a joint normal distribution for (Q_1, Q_2) with correlation 1. The individual variances are thus approximated by (17).

Table 3. A comparison of the approximating distributions of steady-state performance measures in the unbalanced-overload case with simulation results for the base case with $\lambda_1 = 1.3n$ and $\lambda_2 = 0.9n$.

	$n = 25$		$n = 100$		$n = 400$	
Perf. meas.	Approx.	Sim.	Approx.	Sim.	Approx.	Sim.
$std(Q_s)$	16.6	16.0 ± 0.3	33.2	33.7 ± 1.4	66.3	67.6 ± 2.9
$std(\hat{Q}_s)$	3.32	3.21	3.32	3.37	3.32	3.38
$std(Q_1)$	8.3	8.8 ± 0.1	16.6	17.2 ± 0.7	33.2	33.9 ± 1.4
$std(\hat{Q}_1)$	1.66	1.75	1.66	1.72	1.66	1.7
$std(Q_2)$	8.3	8.6 ± 0.1	16.6	17.1 ± 0.7	33.2	33.9 ± 1.5
$std(\hat{Q}_2)$	1.66	1.73	1.66	1.71	1.66	1.69
$\hat{Q}_1$ *quantiles*						
0.05	−2.72	-2.75 ± 0.06	−2.72	-2.84 ± 0.11	−2.72	-2.72 ± 0.19
0.25	−1.12	-1.27 ± 0.08	−1.12	-1.14 ± 0.03	−1.12	-1.18 ± 0.08
0.75	1.12	1.13 ± 0.08	1.12	1.14 ± 0.08	1.12	1.11 ± 0.08
0.95	2.72	2.97 ± 0.11	2.72	2.82 ± 0.20	2.72	2.92 ± 0.16
$\hat{Q}_2$ *quantiles*						
0.05	−2.72	-2.94 ± 0.14	−2.72	-2.82 ± 0.15	−2.72	-2.68 ± 0.21
0.25	−1.12	-1.18 ± 0.08	−1.12	-1.14 ± 0.04	−1.12	-1.17 ± 0.06
0.75	1.12	1.18 ± 0.07	1.12	1.14 ± 0.09	1.12	1.11 ± 0.08
0.95	2.72	2.90 ± 0.10	2.72	2.80 ± 0.20	2.72	2.91 ± 0.15
Centered D quantiles						
0.05	−17.4	-13.4 ± 0.7	−18.4	-16.6 ± 0.6	−19.5	-18.2 ± 0.6
0.25	−7.4	-6.0 ± 0.0	−8.4	-7.6 ± 0.6	−8.5	-8.0 ± 0.0
0.75	−1.4	-0.8 ± 0.6	−1.4	-1.0 ± 0.1	−1.4	-1.0 ± 0.0
0.95	0.5	5.0 ± 1.8	0.5	1.0 ± 0.1	0.5	1.0 ± 0.0

Comparison with simulation. We now compare the approximating normal steady-state distributions to simulation results. We again consider the base case in Table 1 with $\lambda_1 = 1.3n$ and $\lambda_2 = 0.9n$. The results are given in Table 3.

We give the standard deviations of the total queue length $Q_s = Q_1 + Q_2$, as well as the two queues. As before, we treat both the actual values and the scaled values, but now we are scaling in diffusion scale (dividing by $\sqrt{n}$ after subtracting the order-$O(n)$ mean), as in (4), so that we will be substantiating the stochastic process limit in (5). To further substantiate both the stochastic process limit and the normal approximations, we also give the quantiles of the scaled queue lengths $\hat{Q}_1$ and $\hat{Q}_2$. To save space, we omit the confidence intervals for the scaled standard deviations; these can be computed from those of the actual queues by dividing the half-widths by $\sqrt{n}$.

We also give the quantiles for the centered steady-state queue difference $\tilde{D} \equiv D - E[D]$. (Table 1 already showed that the approximation for the mean $E[D]$ is accurate for $n \geqslant 100$.) The approximate distribution of D is obtained from the QBD FTSP. The quantiles of the distribution of $\tilde{D}$ pose a problem because D is integer valued. We thus calculate a linear interpolation of two values. For example, for the 0.05 quantile, we took the largest value d_0 such that $P(\tilde{D} \leqslant d_0) < 0.05$, and linearly interpolate this value with the smallest value d_1 such that $P(\tilde{D} \leqslant d_1) > 0.05$. The linear interpolation becomes just the weighted average of the two values d_0 and d_1. As in Table 1, $\tilde{D}$ is not scaled by any division.

The exact asymptotic distribution. In fact, we have established an FCLT in Perry and Whitt (2011c) that yields the exact asymptotic steady-state distribution of $(Q_1, Q_2, Z_{1,2})$. Consistent with above, the distribution is multivariate normal, but the variances and covariances are different in general; see Corollary 4.1 of Perry and Whitt (2011c). The exact asymptotic results show that there is another term, but it tends to be small. Interestingly, this second term has a contribution from the asymptotic variance of the FTSP D_t. Overall, the FCLT provides strong support for the elementary approximations in (17).

As should be expected, our heuristic OU approximation deteriorates as the difference between θ_1 and θ_2 grows. In extreme cases it might be safer to use the exact diffusion limits (which are harder to analyze), especially if the full distribution of the diffusion approximation is desired. However, simulation experiments show that even if very large differences between the abandonment and service rates of the two classes hold (unlikely in applications), and in addition the system is only lightly overloaded, the heuristic OU process still provides surprisingly accurate steady-state approximations for the variance terms; see EC.6.2 in the authors' homepages, where a very extreme case, having θ_1 ten times larger than θ_2 and a lightly overloaded system, is considered.

8. Conclusions

In this paper we have developed the AP and applied it to describe (i) the transient behavior of the X model during an overload incident and (ii) greatly improve the quality of the steady-state approximation, improving the approximation of the mean queue lengths and obtaining an approximation for the full joint distribution of the queue lengths.

Many open problems remain. First, it remains to develop corresponding performance approximations for the X model with nonexponential distributions. Second, the whole discussion was limited to the overloaded two-class-two-pool X-model setting, but the control and the results should be extended to other MSHT regimes and more complex systems, as in Gurvich and Whitt (2009a, b, 2010). For applications to modern call centers, we would want the two service systems to be more general than the I models considered here. Also, we would like to consider sharing among more than two service systems. The QR-T and FQR-T controls extend quite naturally to more complex systems, but our mathematical analysis, both here and in our other papers, evidently does not extend so easily. Such extensions remain a topic for future research.

Acknowledgments

This research began while the first author was completing his Ph.D. in the Department of Industrial Engineering and Operations Research at Columbia University and was completed while he held a postdoctoral fellowship at Centrum Wiskunde & Informatica (CWI) in Amsterdam. This research was partly supported by NSF grants DMI-0457095 and CMMI 0948190.

References

Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* **16**(6) 665–688.

Bramson, M. 1998. State space collapse with applications to heavy-traffic limits for multiclass queueing networks. *Queueing Systems* **30**(1–2) 89–148.

Coffman, E. G., A. A. Puhalskii, M. I. Reiman. 1995. Polling systems with zero switchover times: A heavy-traffic averaging principle. *Ann. Appl. Probab.* **5**(3) 681–719.

Courtois, P. 1977. *Decomposibility*. Academic Press, New York.

Dai, J. G., T. Tezcan. 2011. State space collapse in many server diffusion limits of parallel server systems. *Math. Oper. Res.* **36**(2) 271–320.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.

Garnett, O., A. Mandelbaum. 2000. An introduction to skill-based routing and its operational complexities. Unpublished manuscript, Technion, Haifa, Israel. http://iew3.technion.ac.il/serveng.

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

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

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

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

Hunt, P. J., T. G. Kurtz. 1994. Large loss networks. *Stochastic Processes Their Appl.* **53**(2) 363–378.

Latouche, G., V. Ramaswami. 1999. *Introduction to Matrix Analytic Methods in Stochastic Modelling*. SIAM and ASA, Philadelphia.

Perry, O., W. Whitt. 2009. Responding to unexpected overloads in large-scale service systems. *Management Sci.* **58**(8) 1353–1367.

Perry, O., W. Whitt. 2011a. An ODE for an overloaded X model involving a stochastic averaging principle. *Stochastic Systems*. Forthcoming.

Perry, O., W. Whitt. 2011b. A fluid limit for an overloaded X model via an averaging principle. Working paper, Columbia University, New York. http://www.columbia.edu/~ww2040/allpapers.html.

Perry, O., W. Whitt. 2011c. Diffusion approximation for an overloaded X model via an averaging principle. Working paper, Columbia University, New York. http://www.columbia.edu/~ww2040/allpapers.html.

Whitt, W. 2002. *Stochastic-Process Limits*. Springer, New York.

Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* **50**(10) 1449–1461.