MANAGEMENT SCIENCE
DOI 10.1287/mnsc.1090.1025ec

e-companion

ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"Responding to Unexpected Overloads in Large-Scale Service Systems" by Ohad Perry and Ward Whitt, *Management Science*, DOI 10.1287/mnsc.1090.1025.

e-Companion

In this online e-companion we present additional material supplementing the main paper. The topics are ordered as they arise in the paper. In §EC.1 we discuss the way the transient distribution approaches its steady-state limit, both at the beginning and the end of an overload incident. In §EC.2 we provide additional discussion about the FQR and FQR-T controls, supplementing §4. In §EC.3 we present additional details about the optimal solution for the deterministic fluid model during the overload, supplementing §5. Finally, in §EC.5 we present additional simulation results about the performance of the control. In §EC.5.1 we present a table of detailed simulation results supporting Figure 5. In §EC.5.2 we present additional simulation results about the performance of FQR-T under normal loading. We perform a sensitivity analysis for the thresholds there.

EC.1. Time To Reach Steady State

An important aspect of our QR-T and FQR-T controls is the transient behavior of the system. When the overload incident occurs, the system must shift from steady state under normal loading to steady state under the overload. Afterwards, at the end of the overload period, there is a recovery period, during which the system shifts back to the original steady state. From analysis and extensive simulations, we conclude that these two transient periods do not dominate, so that it is possible to use steady-state analysis as a reasonable approximation. In this section, we provide some supporting simulation results and discuss the supporting mathematical results.

EC.1.1. Simulation Experiments

We start by doing a simulation experiment of an overload incident, including all five regimes: (i) steady state before the overload, (ii) transition to new steady state at the beginning of the overload, (iii) new steady state under the overload, (iv) recovery period and (v) original steady state again after the overload.

Our example is based on Example 1 in the main paper and the associated typical overload incident described at the end of §3. We assume that there is an overload incident that lasts 5 hours when the mean service times are 5 minutes. Given that we measure time in units of mean service

times, the overload incident lasts 60 time units. Thus, we simulate the system over the time interval $[0,150]$, and have the overload begin at time 80 and end at time 140. Thus, the initial transient begins at time 80, while the recovery period begins at time 140.

We consider a large system with $n=400$ agents in each pool. For the normal loading, we let $\lambda_1 = \lambda_2 = 380$; for the overload during $[80,140]$, we let $\lambda_1 = 520$, while $\lambda_1 = 380$ as before. As in Example 1, we let the mean service time for customers served by designated agents be $\mu_{1,1}^{-1} = \mu_{2,2}^{-1} = 1.0$, while the mean service time for customers served by agents from the other pool is $\mu_{1,2}^{-1} = \mu_{2,1}^{-1} = 1.25$. We let customers abandon at rate $\theta_1 = \theta_2 = 0.4$.

Since class 1 experiences the overload, we will have pool 2 helping class 1 during the overload incident. Typical sample paths of the processes $Z_{1,2}(t)$ and $Q_1(t)$ generated by simulation are shown in Figures EC.1 and EC.2 below. A dotted horizontal line depicts the steady-state fluid approximation during the overload. We do not show the other processes. From corresponding plots of $Q_2(t)$ and $Q_1(t)$, it is evident that they move together during the overload, reflecting state-space collapse, but they move independently during normal loading. From the displayed sample path, we

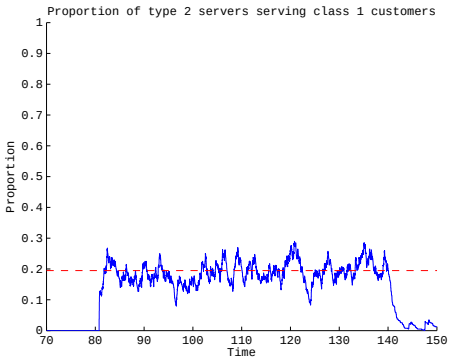

Figure EC.1 Sample path of $Z_{1,2}(t)/400$ for FQR-T, with overload incident in $[80,140]$ for $n=400$.

Figure EC.2 Sample path of $Q_1(t)$ for FQR-T, with overload incident in $[80,140]$ for $n=400$.

see that the system indeed reaches a new steady state after a few mean service times, as claimed in §§2-3.

To elaborate, we also show corresponding sample paths in Figures EC.3 and EC.4 with $n=100$

agents in each pool. One important observation is that in both systems ($n = 100$ and $n = 400$) it takes less than 3 time units for the queues to hit their fluid value, denoted by the dotted horizontal lines. The recovery time, after the overload incident has ended, is also very short, and is about 2 time units for the queues in both systems.

The story is different for the $Z_{1,2}(t)$ process. To make the connection between the two cases clear, we present the **proportion** of class-1 customers in pool 2 instead of the actual number, i.e., we show $Z_{1,2}(t)/n$ in Figures EC.1 and EC.3. First, when the overload begins at time 80, it takes some time until the queues hit the threshold $\kappa_{1,2}$. That is the reason why $Z_{1,2}(t)$ starts growing a bit after time 80. It is interesting to see how our choice of the thresholds influences this delay. Recall that we choose the thresholds to be of order of size less than $O(n)$ but greater than $O(\sqrt{n})$; see §6 for more details. In these simulations, we took $\kappa_{i,j} = 20$ for $n = 400$ and $\kappa_{i,j} = 10$ for $n = 100$. This explains why in the $n = 400$ system it takes less time for $Z_{1,2}(t)$ to start increasing than in the $n = 100$ system: The thresholds are relatively smaller for the bigger system.

We also observe a difference between the two systems after the arrival rates return to normal at time 140. At this time, the $Z_{1,2}(t)$ processes start decreasing immediately and in a very fast rate. But now, service-pool 2 stops serving class-1 customers faster in the small system. Let $T_{1,2}$ be the time it takes for pool 2 to stop serving all class-1 customers after the end of the overload incident (after 140 in our example). As an approximation, we have

$$E[T_{1,2}] \approx \sum_{j=1}^{r} \frac{1}{j \cdot \mu_{1,2}},$$

where $r \equiv Z_{1,2}(140)$. Hence, the larger $Z_{1,2}(140)$ is, the longer it takes $Z_{1,2}(t)$ (or equivalently, $Z_{1,2}(t)/n$) to reach zero after the arrival rates shift back to normal. Yet, in both cases $Z_{1,2}(t)/n$ drops below 0.1 in about 2 time units, so that the total service rate in service-pool 2 is greater than λ_2 in 2 time units after the shift. In summary, we see that the transient period is relatively short, and a steady-state analysis is reasonable to apply.

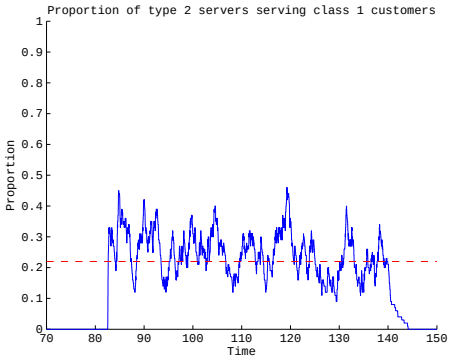

Figure EC.3 **Sample path of $Z_{1,2}(t)/100$ for FQR-T, with overload incident in $[80, 140]$ for $n = 100$.**

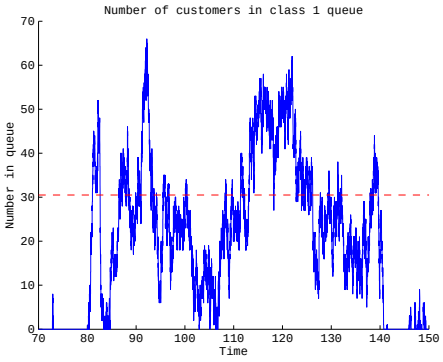

Figure EC.4 **Sample path of $Q_1(t)$ for FQR-T, with overload incident in $[80, 140]$ for $n = 100$.**

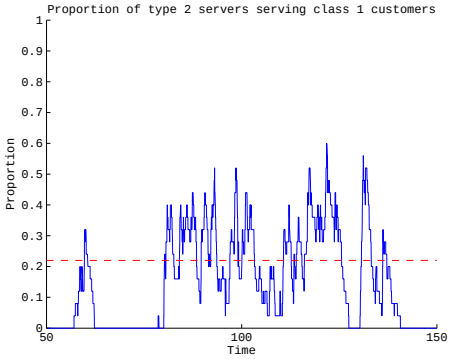

Figure EC.5 **Sample path of $Z_{1,2}(t)/25$ for FQR-T, with overload incident in $[80, 140]$ for $n = 25$.**

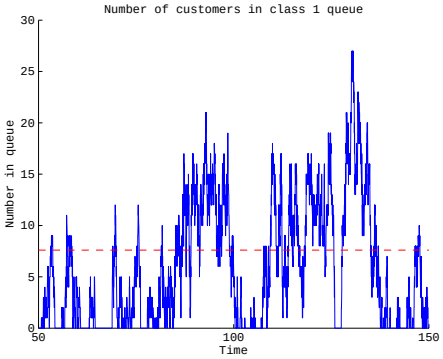

Figure EC.6 **Sample path of $Q_1(t)$ for FQR-T, with overload incident in $[80, 140]$ for $n = 25$.**

EC.1.2. Mathematical Analysis

We now provide further support. We first review mathematical analysis of the $M/M/n+M$ model; we next contrast with single-server models; afterwards we discuss implications for our X model

The M/M/n+M model. Consider the $M/M/n + M$ model with arrival rate λ, service rate μ and abandonment rate θ. First, it is useful to consider the special case in which $\theta = \mu$; then the number in system is distributed the same as in an $M/M/\infty$ system with service rate $\theta = \mu$. Thus the number in system at time t has a Poisson distribution for each fixed initial state. An explicit expression for the mean $m(t)$ at time t, starting empty, is given in (20) of Eick et al. (1993). More generally, the mean $m(t)$ satisfies an ordinary differential equation (ODE); see Corollary 4

of Eick et al. (1993). These results show that $m(t)$ and the entire distribution reaches steady state approximately at time c/μ, some constant c times the mean service time $1/\mu$. The constant c depends on our criterion; the critical time constant is $1/\mu$, a mean service time.

For the more general overloaded $M/M/n+M$ model (without assuming that $\theta = \mu$), it is helpful to consider the deterministic fluid approximation in Whitt (2004). Formula (2.17) there shows that the fluid approximation for the number in queue, $q(t)$, starting with all the servers busy, again evolves as the $M/M/\infty$ ODE, but with arrival rate $\lambda - n\mu$ and service rate θ. That implies that the fluid queue content (approximating the number in queue), starting from all servers busy but no queue, reaches steady state approximately at time c/θ, some constant c times the mean abandonment time $1/\theta$. That too will be approximately c/μ provided that θ is not too different from μ. The critical time constant here is $1/\theta$, a mean time to abandon.

To illustrate this mathematical analysis, we do a simulation of the $M/M/n+M$ model. We base our example here on Example 1 in §1. In that example, the service rates in both pools are $\mu_{i,i} = 1$, the abandonment rates are $\theta_i = 0.4$ and the number of agents in each pool is 100. In this example the arrival rates changed at some instant from $(\lambda_1, \lambda_2) = (90, 90)$ to $(\lambda_1, \lambda_2) = (130, 90)$. We show what happens if class 1 receives no help from service-pool 2. Then the class-1 queue behaves like an $M/M/100 + M$ queue. Figure EC.7 depicts a simulated sample path of an $M/M/100 + M$ queue, when the system is initialized empty at time 0. The average steady-state queue length in the overload incident is about 75, and it can be seen that this steady-state value is reached within about 4 time units when the system is initialized empty. (Time is measured in units of mean service times). If we assume, as in our example above, that the system was operating before the arrival rates changed, then most of the agents were probably busy, and the time to reach the new steady state is about 2 time units (two mean service times).

Single-Server models. In §2 we stated that the number in system tends to approach steady state more quickly in many-server queues with abandonment than in single-server queues without abandonment. We should begin with a qualification: Slow approach to steady state occurs for single-server systems without abandonment when the system is heavily loaded. For single-server

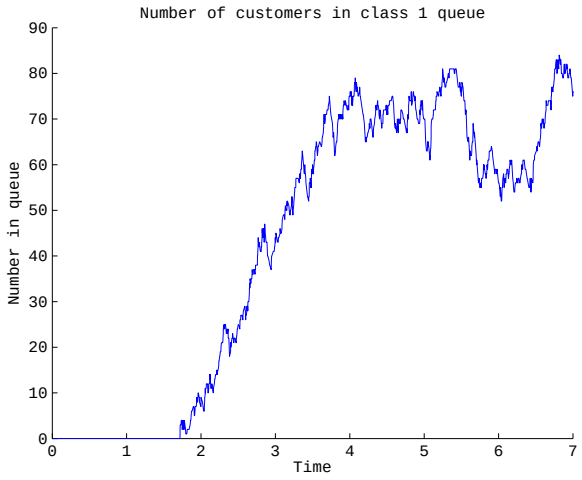

Figure EC.7 **Time to reach steady state when** $m_1 = m_2 = 100$, $\lambda_1 = 130$, $\lambda_2 = 90$, $\mu_{1,1} = \mu_{2,2} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.8$ **and** $\theta_1 = \theta_2 = 0.4$.

queues, we refer to Section III.7.3 of Cohen (1982) on the relaxation time. Sections 4.6 and 5.1 of Whitt (1989) gives conventional heavy-traffic approximations (when $\rho \uparrow 1$ with n fixed, where $\rho \equiv \lambda/n\mu$ is the traffic intensity) for the time required for the mean number in system to reach steady state in the general $G/G/n$ model with fixed n and without customer abandonment. The time required to reach steady state is approximately $c/(1-\rho)^2$ mean service times, where c is a constant depending on the number of servers, n, the variability of the arrival and service processes (quantified explicitly) and again the criterion. Clearly the time to reach steady state can be quite long when ρ is high.

The X model. For our X model, there are two implications of the $M/M/n+M$ analysis above: First, when the overload incident begins, the queue length should be negligible, so that the fluid content in a newly overloaded queue will grow approximately linearly at rate $\lambda - n\mu$, because the opposite force $\theta q(t)$ will be small, since $q(t)$ is initially small. That means that the threshold will be quickly passed if there is a significant unbalanced overload.

For our more complicated X model with the QR-T control, after the threshold has been exceeded, the theoretical analysis for the $M/M/n+M$ model above provides a rough heuristic analysis indicating what should happen, but the actual evolution still depends on the state of the six-dimensional Markov chain $(Q_i(t), Z_{i,j}(t); i = 1, 2; j = 1, 2)$. Thus we rely on simulation to confirm

that the actual behavior is indeed similar to what occurs in these simple many-server models. We remark that the state-space collapse discussed in the next subsection indicates that $(Q_1(t), Q_2(t))$ should evolve approximately as a one-dimensional process, suggesting that the analysis above should not be too far off when the service rates $\mu_{i,j}$ do not differ greatly.

EC.2. More on FQR

In this section we present additional background on FQR; for more, see Gurvich and Whitt (2007a,b,c,d). We first illustrate the state-space collapse (SSC), which is the topic of Gurvich and Whitt (2007c). The conditions for SSC are satisfied if either the service rates only depend upon the customer class or the service rates only depend upon the agent pool. To illustrate, suppose that the service rates are independent of both class and pool, with $\mu_{1,1} = \mu_{1,2} = \mu_{2,1} = \mu_{2,2} = 1.0$. Figure EC.8 shows the plots of typical sample paths of the two queue-length processes when $\lambda_1 = \lambda_2 = m_1 = m_2 = 100$ and $\theta_1 = \theta_2 = 0.2$. From Figure EC.8, we can clearly see the SSC.

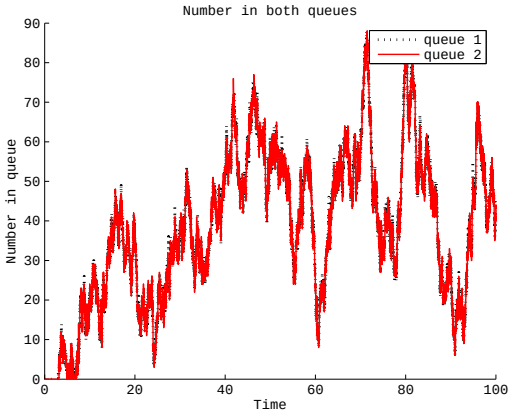

Figure EC.8 State-Space Collapse

We observed that, with FQR, it is possible to choose the ratio parameter r (or, equivalently, the queue proportions p_i) in order to determine the optimal level of staffing to achieve desired service-level differentiation. For example, under normal loading, our goal may be to choose staffing levels as small as possible subject to having 80% of class-1 customers wait less than 20 seconds, while 80% of class-2 customers wait less than 60 seconds. To see how this can be done with FQR, let T_i be

the class-i delay target (e.g., $T_1 = 0.033$ and $T_2 = 0.100$ for 20 seconds and 60 seconds if the mean service times are 10 minutes); let W_i be the class-i waiting time before starting service; let p_i be the queue proportion determined by r. As explained in Gurvich and Whitt (2007a), the following string of approximations show how the individual class-i performance targets $P(W_i > T_i) \le \alpha$, for both i, can be reduced into a single-class single-pool performance target $P(W > T) \le \alpha$ for an appropriate choice of the queue proportions p_i and the aggregate target T:

$$\begin{aligned} P(W_i > T_i) &\approx P(Q_i > \lambda_i T_i) \approx P\left(p_i Q_\Sigma > \lambda_i T_i\right) \approx P\left(Q_\Sigma > \sum_{k=1}^{2} \lambda_k T_k\right) \\ &\approx P\left(\lambda W > \sum_{k=1}^{2} \lambda_k T_k\right) \approx P(W > T) \le \alpha \ , \end{aligned} \tag{EC.1}$$

where we define $p_i \equiv \lambda_i T_i/(\lambda_1 T_1 + \lambda_2 T_2)$, $\lambda \equiv \lambda_1 + \lambda_2$ and $T \equiv (\lambda_1 T_1 + \lambda_2 T_2)/(\lambda_1 + \lambda_2)$. The first approximation in (EC.1) follows by a heavy-traffic generalization of Little's law, establishing that the steady-state queue-length and waiting-time random variables are related approximately by $Q_i \approx \lambda_i W_i$. The second approximation in (EC.1) is due to SSC: $Q_i \approx p_i Q_\Sigma$. The third approximation is obtained by choosing p_i as specified above. The fourth approximation in (EC.1) follows from the heavy-traffic generalization of Little's law once again, for the entire system: $Q_\Sigma \approx \lambda W$ for λ as defined above, where W is the waiting time for an arbitrary customer. The fifth and final approximation follows by the appropriate definition of the aggregate target T, as defined above. With this reduction, we can determine the overall staffing by using elementary established methods for the single-class single-pool model. That is, we choose the total number of agents, m, so that $P(W > T) \le \alpha$ in the $M/M/m + M$ model. We then let $m_i = p_i m$. From (EC.1) and the fact that $r = p_1/(1 - p_1)$, we see that the required ratio is

$$r = \frac{p_1}{1 - p_1} = \frac{\lambda_1 T_1}{\lambda_2 T_2}. \tag{EC.2}$$

For the X model (and more generally), Theorem 4.1 of Gurvich and Whitt Gurvich and Whitt (2007d) shows that, if the service rates only depend on the service pool or the class (but not both), then FQR is asymptotically optimal to minimize linear staffing costs subject to service-level constraints, as above, in the QED many-server heavy-traffic regime.

As was shown in §4.1, with inefficient sharing FQR without the thresholds we add in FQR-T can cause the queues in the general X-model system to explode when there is no abandonment, because of the inefficient sharing. We now show that there is also serious performance degradation when we include customer abandonment. We use the same example as in §4.1, only adding abandonments with rates $\theta_1 = \theta_2 = 0.2$. As before, there are 100 agents in each pool. The arrival rates are $\lambda_1 = \lambda_2 = 99$ and the service rates are $\mu_{1,1} = \mu_{2,2} = 1$ and $\mu_{1,2} = \mu_{2,1} = 0.8$. To describe the performance degradation, we compare the performance to the no-sharing case. When there is no sharing, 2% of the customers abandon, the mean queue length is 10 and the mean conditional waiting time given that the customer is served is 0.10. On the other hand, for FQR with $r = 1$, again about 39% of the agents are busy serving customers from the other class. That reduces the effective service rate for each class from 100 to 92.2. As a consequence, about 7% of the customers abandon, the mean queue length is 34 and the average conditional waiting time given that the customer is served is 0.35.

Figures EC.9 and EC.10 show the sample paths of the number of agents in pool 1 helping class-2 customers, and the class-1 queue, respectively. Due to the symmetry of the system in our example, the $Z_{2,1}$ and Q_2 figures are very similar, and the fluid approximations for both queues and $Z_{i,j}$'s are equal.

Figure EC.9 Sample path of $Z_{2,1}(t)/100$ for FQR with $r = 1$ and abandonments at rate $\theta_1 = \theta_2 = 0.2$.

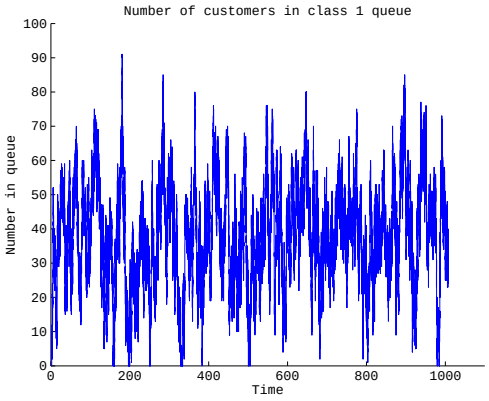

Figure EC.10 Sample path of $Q_1(t)$ for FQR with $r = 1$ and abandonments at rate $\theta_1 = \theta_2 = 0.2$.

In contrast, to illustrate how FQR-T performs, we consider the same example: Example 2 with abandonments at rate $\theta_i = 0.2$. We let $r_{1,2} = r_{2,1} = 1$, so that there is no change from FQR above, and we let the thresholds be $\kappa_{1,2} = \kappa_{2,1} = 10$. The results of a simulation experiment are shown in Figures EC.11 and EC.12. Numerical values were given in §4.2. The performance is greatly

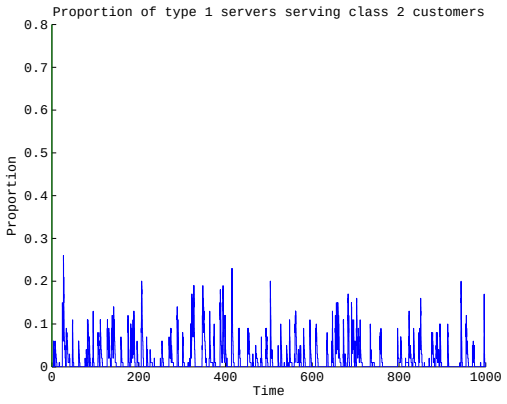

Figure EC.11 **Sample path of $Z_{2,1}(t)/100$ with FQR-T, $r = 1$.**

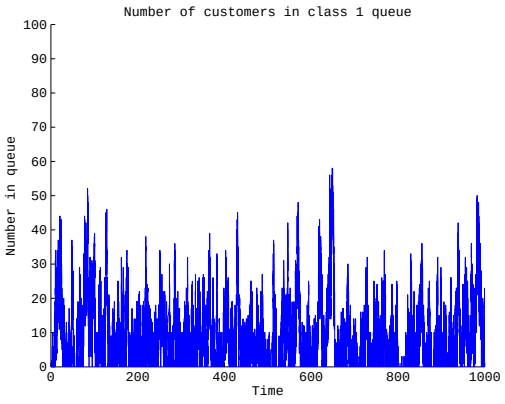

Figure EC.12 **Sample path of $Q_1(t)$ with FQR-T, $r = 1$.**

improved with FQR-T.

EC.3. Optimal Solution for the Fluid Model

In this section we provide additional material supplementing §5.

EC.3.1. Proof of Proposition 3.

The representation is an immediate consequence of Proposition 2. Since $C_{1,2}(Z_{1,2})$ is the composition of a strictly convex function and a linear function, it is a strictly convex function of $Z_{1,2}$; e.g., p. 38 of Rockafellar (1970); similarly for $C_{2,1}(Z_{2,1})$. To establish the convexity of C_c, first assume that C is differentiable. It suffices to show that the derivative of C_c with respect to $Z_{1,2} - Z_{2,1}$, denoted by C_c', is nondecreasing. Existence and monotonicity of the derivative C_c' away from the boundary point $Z_{1,2} - Z_{2,1} = 0$ follows from the differentiability and convexity of $C_{1,2}$ and $C_{2,1}$, assuming that C is differentiable and convex. However, even if C is differentiable, the derivative of C_c need not exist at $Z_{1,2} - Z_{2,1} = 0$. It suffices to show that the left derivative is less than the right derivative at this point. The right derivative of C_c at 0, denoted by $C_c'^{+}(0)$, coincides

with the derivative $C'_{1,2}(0)$, while the left derivative of C_c at 0, denoted by $C_c'^{-}(0)$, coincides with $-C'_{2,1}(0)$. Let C'_i denote the partial derivative of C with respect to its i^{th} coordinate at the argument $(q_1-(s_1\mu_{1,1}/\theta_1), q_2-(s_2\mu_{2,2}/\theta_2))$, which is positive because C is increasing. Then observe that

$$C'_{1,2}(0) = -C'_1\left(\frac{\mu_{1,2}}{\theta_1}\right) + C'_2\left(\frac{\mu_{2,2}}{\theta_2}\right) \quad \text{and} \quad -C'_{2,1}(0) = -C'_1\left(\frac{\mu_{1,1}}{\theta_1}\right) + C'_2\left(\frac{\mu_{2,1}}{\theta_2}\right)$$

Hence, $C'_{1,2}(0) \geq -C'_{2,1}(0)$, so that $C_c'^{+}(0) > C_c'^{-}(0)$ if the two inequalities in (1) hold. These relations can be extended to non-differentiable functions C by working with left and right derivatives. ∎

EC.3.2. Optimal Values Beyond the Boundaries

It is natural to have the cost function C be smooth, in which case the optimal solution can be found by simple calculus. The following result concludes that, if the optimal solution found by calculus falls outside the feasible set, then the actual optimum value is obtained at the nearest boundary point. Let $a \wedge b \equiv \min\{a,b\}$ and $a \vee b \equiv \max\{a,b\}$. We omit the proof, which is a standard convexity result.

PROPOSITION EC.1. (optimal values beyond the boundaries) *Let $\bar{Z}_{1,2}$ and $\bar{Z}_{2,1}$ be the values of $Z_{1,2}$ and $Z_{2,1}$ yielding minimum values of $C_{1,2}$ and $C_{2,1}$ in* (9)*, and let $\hat{Z}_{1,2}$ and $\hat{Z}_{2,1}$ be the corresponding values yielding the minima ignoring the constraints in Proposition 3. Then $\bar{Z}_{1,2} = \hat{Z}_{1,2} \vee 0 \wedge m_2$, $\bar{Z}_{2,1} = \hat{Z}_{2,1} \vee 0 \wedge m_1$ and $(Z^*_{1,2}, Z^*_{2,1})$ can assume only two possible values: $(\bar{Z}_{1,2}, 0)$ or $(0, \bar{Z}_{2,1})$.*

EC.3.3. The Relation between r and Z

In §5.2 we observed that there is a one-to-one correspondence between the queue ratio $r \equiv Q_1/Q_2$ and the real variable $Z_{1,2} - Z_{2,1}$ used to specify the optimization problem in Proposition 3. That implies that there is a one-to-one correspondence between the fixed-agent-allocation optimization problem (choosing $Z_{1,2}$ and $Z_{2,1}$) and the queue-ratio control problem (choosing a state-dependent queue-ratio r) in the fluid-model context.

PROPOSITION EC.2. (relating r and $Z_{1,2} - Z_{2,1}$) *For any given arrival-rate vector* (λ_1, λ_2) *or initial state* (q_1, s_1, q_2, s_2) *(without sharing), the queue ratio* $r \equiv Q_1/Q_2$ *is a strictly decreasing differentiable function of* $Z_{1,2} - Z_{2,1}$*, denoted by* ϕ*, as* $Z_{1,2} - Z_{2,1}$ *varies over its allowed domain in Proposition 3. Thus, the function* ϕ *has a unique inverse* ϕ^{-1} *and there exists a unique optimal* $r^* \equiv r^*(q_1, s_1, q_2, s_2)$*, which is characterized by*

$$r^* = \phi^{-1}(Z^*_{1,2} - Z^*_{2,1}), \tag{EC.3}$$

where both r^* *and* $Z^*_{1,2} - Z^*_{2,1}$ *are understood to be functions of the initial state* (q_1, s_1, q_2, s_2)*. Moreover, there are two thresholds* $\eta_{1,2} > \eta_{2,1}$ *such that we want one-way sharing with pool* 2 *helping class* 1 *if* $r > \eta_{1,2}$*, in which case we let* $r_{1,2} = r^*$*; we want one-way sharing with pool* 1 *helping class* 2 *if* $r < \eta_{2,1}$*, in which case we let* $r_{2,1} = r^*$*; and we want no sharing at all if* $\eta_{2,1} \le r \le \eta_{1,2}$*. The thresholds are obtained from the thresholds* $\zeta_{1,2}$ *and* $\zeta_{2,1}$ *in Corollary* 1 *by* $\eta_{1,2} = \phi^{-1}(\zeta_{1,2})$ *and* $\eta_{2,1} = \phi^{-1}(\zeta_{2,1})$*.*

Proof. By (9), when pool 2 helps class 1, Q_1 is a strictly decreasing differentiable function of $Z_{1,2}$ and while Q_2 is a strictly increasing differentiable function of $Z_{1,2}$. On the other hand, when pool 1 helps class 2, Q_1 is a strictly increasing differentiable function of $Z_{2,1}$ and while Q_2 is a strictly decreasing differentiable function of $Z_{2,1}$. Thus $r \equiv Q_1/Q_2$ is a strictly decreasing differentiable function of $Z_{1,2} - Z_{2,1}$ over its domain, as claimed. ∎

EC.3.4. Constant Weighted Queue Length

We now complete Proposition 4 by exhibiting the result for pool 1 helping class 2.

PROPOSITION EC.3. (constant weighted queue lengths with pool 1 helping class 2) *Let*

$$a_{2,1} \equiv \frac{\mu_{2,1}\theta_1}{\mu_{1,1}\theta_2} \quad \textit{and} \quad \tilde{a}_{2,1} \equiv \frac{\mu_{2,1}}{\mu_{1,1}}. \tag{EC.4}$$

Consider any initial state (λ_1, λ_2)*, or equivalently* (q_1, s_1, q_2, s_2)*, with* $s_2 = 0$*. Let*

$$w_{2,1} \equiv a_{2,1}\left(\frac{\lambda_1 - m_1\mu_{1,1}}{\theta_1}\right) + \left(\frac{\lambda_2 - m_2\mu_{2,2}}{\theta_2}\right) = a_{2,1}\left(q_1 - \frac{s_1\mu_{1,1}}{\theta_1}\right) + q_2. \tag{EC.5}$$

Then

$$a_{2,1}\left(Q_1(Z_{1,2}) - \frac{S_1(Z_{2,1})\mu_{1,1}}{\theta_1}\right) + Q_2(Z_{2,1}) = w_{2,1} \tag{EC.6}$$

for all $Z_{2,1}$ *with* $0 \le Z_{2,1} \le m_1$.

Just as with Proposition 4, Proposition EC.3 implies that the locus of all nonnegative queue-length vectors $(Q_1, Q_2) \equiv (Q_1(Z_{2,1}), Q_2(Z_{2,1}))$ associated with initial state (λ_1, λ_2), or equivalently (q_1, s_1, q_2, s_2), with $s_2 = 0$, is on the line $\{(Q_1, Q_2) : a_{2,1}Q_1 + Q_2 = w_{2,1}\}$ in the nonnegative quadrant. Thus, for any nonnegative constant $w_{2,1}$, the optimal queue-length vector (Q_1^*, Q_2^*) and the optimal queue-ratio $r_{2,1}^* \equiv Q_1^*/Q_2^*$ restricted to one-way sharing ($Z_{1,2} = 0$) are the same for all initial states (q_1, s_1, q_2, s_2) with $s_2 = 0$ satisfying *(13)* and $q_2 \ge Q_2^*$. Moreover, $a_{2,1}Q_1^* + Q_2^* = w_{2,1}$. That same optimal queue-length vector and optimal queue ratio holds for all arrival pairs (λ_1, λ_2) where $s_2 = 0$, $Z_{1,2} = 0$ and

$$\lambda_1 + \tilde{a}_{2,1}\lambda_2 = \tilde{w}_{2,1} \equiv \frac{\theta_1\theta_2 w_{2,1} + a_{2,1}\theta_2 m_1\mu_{1,1} + \theta_1 m_2\mu_{2,2}}{a_{2,1}\theta_2}. \tag{EC.7}$$

EC.4. Structured Separable Cost Functions

At the end of §5.3, we observed that we can obtain explicit analytical expressions for the optimal ratio control if we impose additional structure on our cost function. We give the main results in this section and provide supporting details in the next section.

EC.4.1. Main Results

We first assume that C is separable, i.e., $C(Q_1, Q_2) = C_1(Q_1) + C_2(Q_2)$, where each component cost function C_i is strictly convex, strictly increasing and twice differentiable. We start by assuming that the derivatives C_i' are strictly increasing, so that their inverses exist. Let $\Psi(Q_1) \equiv C_1'(Q_1)$ and let Ψ^{-1} be its inverse. Then one of the following relations between the queue lengths should hold, when we choose the one that minimizes the cost:

$$Q_1 = \Psi^{-1}(a_{1,2}C_2'(Q_2)) \quad \text{or} \quad Q_1 = \Psi^{-1}(a_{2,1}C_2'(Q_2)), \tag{EC.8}$$

for $a_{1,2}$ defined in (12) and $a_{2,1}$ defined in §EC.3.4. If C_1' is not strictly increasing, then we work with the left-continuous inverse of Ψ defined by $\Psi^{\leftarrow} \equiv \{x : \Psi(x) \ge y\}$.

Power functions. If the cost functions C_i are simple power functions, i.e., $C_i(Q_i) \equiv c_i Q_i^{n_i}$ for $i = 1, 2$, then we have that either $Q_1^* = r_{1,2}^* Q_2^{*(n_2-1)/(n_1-1)}$ or $Q_1^* = r_{2,1}^* Q_2^{*(n_1-1)/(n_2-1)}$, where

$$r_{1,2}^* \equiv \sqrt[n_1-1]{a_{1,2}c_2 n_2/c_1 n_1} \qquad \text{and} \qquad r_{2,1}^* \equiv \sqrt[n_1-1]{a_{2,1}c_2 n_2/c_1 n_1}. \tag{EC.9}$$

When $n_1 = n_2$, Q_1^*/Q_2^* is a fixed queue ratio, either $r_{1,2}^*$ or $r_{2,1}^*$ for $r_{i,j}^*$ as in (EC.9). Thus, we need only to decide which way we should share, and then use FQR-T with the appropriate $r_{i,j}^*$; i.e., we are in the setting of Figure 4 with constant ratios for which we have explicit expressions.

Quadratic functions. In practice it may be difficult to actually specify an appropriate cost function. Thus, for practical application we suggest quadratic functions: $C_i(Q_i) \equiv c_i Q_i^2 + b_i Q_i + a_i$ for $i = 1, 2$. These functions might be obtained by performing an approximation (e.g., via Taylor series approximation to an analytical expression or least squares fit to data). In this case, we have either

$$Q_1^* - r_{1,2}^* Q_2^* = k_{1,2}^*, \qquad \text{or} \qquad Q_1^* - r_{2,1}^* Q_2^* = k_{2,1}^*, \tag{EC.10}$$

for

$$r_{1,2}^* \equiv \frac{a_{1,2}c_2}{c_1} = \frac{c_2\mu_{2,2}\theta_1}{c_1\mu_{1,2}\theta_2}, \qquad r_{2,1}^* \equiv \frac{a_{2,1}c_2}{c_1} = \frac{c_2\mu_{2,1}\theta_1}{c_1\mu_{1,1}\theta_2}. \tag{EC.11}$$

and

$$k_{1,2}^* \equiv \frac{a_{1,2}b_2 - b_1}{2c_1}, \qquad k_{2,1}^* \equiv \frac{a_{2,1}b_2 - b_1}{2c_1}. \tag{EC.12}$$

In other words, we keep a fixed-queue ratio centered about a constant $k_{i,j}^*$ instead of zero. That is, we employ new thresholds after sharing has been activated. (The current thresholds $k_{i,j}^*$ are not to be confused with the thresholds $k_{i,j}$ used with the queue-difference processes in (4) to test for the occurrence of overloads. We use $k_{1,2}^*$ only after sharing with pool 2 helping class 1.) From the two formulas in (EC.11), we directly see how these ratio parameters and thresholds should depend on the model parameters. In particular, each ratio is either directly proportional or inversely proportional to each of six model parameters.

Quadratic and linear power functions. A natural simple cost function is the quadratic power function, which is is special case of the general power function with $n_1 = n_2 = 2$ and a special case

of the general quadratic function with $b_1 = b_2 = a_1 = a_2 = 0$. The optimal control then is precisely FQR-T ($k_1^* = k_2^* = 0$), as indicated in Proposition 5. In §EC.4.2 we also discuss the special cases $n_1 = 2, n_2 = 1$ and $n_1 = 1, n_2 = 1$.

FQR-T without cost functions. Clearly, FQR-T could be employed directly without specifying any cost function, using engineering judgment to set the parameters. Even if that is the case, the queue-ratio formulas in (EC.11) and possibly the centering formulas in (EC.12) provide important insight into how the control parameters should depend on the model parameters.

EC.4.2. Supporting Details About Structured Separable Cost Functions

We now supplement §EC.4 by providing more details about the fluid model with a separable cost function. As before, we assume that C is separable, and that each component cost-function C_i is strictly convex, strictly increasing and twice differentiable. We then relax the strictly-increasing assumption, and consider linear functions.

Let $C(Q_1, Q_2)$ be a separable function. We can write C as a function of one variable $Z_{1,2}$ or $Z_{2,1}$, depending on which way the sharing is done.

$$C(Q_1, Q_2) = C_1(Q_1) + C_2(Q_2) \equiv C_{1,(i,j)}(Z_{i,j}) + C_{2,(i,j)}(Z_{i,j}) \equiv C(Z_{i,j}), \tag{EC.13}$$

where

$$C_{1,(1,2)}(Z_{1,2}) \equiv C_1\left(q_1 - \frac{s_1\mu_{1,1}}{\theta_1} - \frac{Z_{1,2}\mu_{1,2}}{\theta_1}\right), \quad C_{2,(1,2)}(Z_{1,2}) \equiv C_2\left(q_2 - \frac{s_2\mu_{2,2}}{\theta_2} + \frac{Z_{1,2}\mu_{2,2}}{\theta_2}\right).$$

and

$$C_{1,(2,1)}(Z_{2,1}) \equiv C_1\left(q_1 - \frac{s_1\mu_{1,1}}{\theta_1} + \frac{Z_{2,1}\mu_{1,1}}{\theta_1}\right), \quad C_{2,(2,1)}(Z_{2,1}) \equiv C_2\left(q_2 - \frac{s_2\mu_{2,2}}{\theta_2} - \frac{Z_{2,1}\mu_{2,1}}{\theta_2}\right).$$

Hence, the optimal $Z_{1,2}$ is achieved when

$$C'(Z_{1,2}) = -C'_{1,(1,2)}(Z_{1,2})\left(\frac{\mu_{1,2}}{\theta_1}\right) + C'_{2,(1,2)}(Z_{1,2})\left(\frac{\mu_{2,2}}{\theta_2}\right) = 0,$$

or equivalently,

$$C'_1(Q_1) = \frac{\mu_{2,2}\theta_1}{\mu_{1,2}\theta_2} C'_2(Q_2) \equiv a_{1,2} C'_2(Q_2). \tag{EC.14}$$

Similarly, the optimal $Z_{2,1}$ is achieved when

$$C_1'(Q_1) = \frac{\mu_{2,1}\theta_1}{\mu_{1,1}\theta_2} C_2'(Q_2) \equiv a_{2,1} C_2'(Q_2). \tag{EC.15}$$

The fact that C_i is strictly convex implies that $C_i'' \geq 0$. If $C_i'' > 0$ then C_i' is strictly increasing, and its inverse function exists. Let $\Psi(Q_1) \equiv C_1'(Q_1)$ and let Ψ^{-1} be its inverse. Then one of the following relations between the queues should hold:

$$\text{either} \quad Q_1 = \Psi^{-1}(a_{1,2} C_2'(Q_2)) \quad \text{or} \quad Q_1 = \Psi^{-1}(a_{2,1} C_2'(Q_2)), \tag{EC.16}$$

where we choose the relation that minimizes the cost-function $C(Q_1, Q_2)$.

If the inverse of Ψ does not exist, then we can work with the left-continuous inverse of Ψ defined by $\Psi^{\leftarrow}(y) \equiv \{x : \Psi(x) \geq y\}$.

We now consider separable cost functions of the form:

$$C(Q_1, Q_2) = c_1 Q_1^{n_1} + c_2 Q_2^{n_2}, \qquad n_1, n_2 \in \mathbb{N}, \tag{EC.17}$$

where each component is a power function. The optimal solution is given in the main paper. We observe that the compatible-ratio condition $r_{1,2} \geq r_{2,1}$ in S 4.2 holds, because

$$\frac{r_{1,2}^*}{r_{2,1}^*} = \sqrt[n_1-1]{\frac{a_{1,2}}{a_{2,1}}} = \sqrt[n_1-1]{\frac{\mu_{1,1}\mu_{2,2}}{\mu_{1,2}\mu_{2,1}}} \geq 1,$$

under the inefficient-sharing condition (2). When $n_1 = n_2$ we get

$$\frac{Q_1^*}{Q_2^*} = r_{1,2}^* \quad \text{or} \quad \frac{Q_1^*}{Q_2^*} = r_{2,1}^*;$$

i.e., it is optimal to keep a fixed-queue ratio. Thus, we need only to decide on which way we should share, and then use FQR-T with the appropriate fixed queue ratio $r_{i,j}^*$.

These results explain why the optimal ratios in our numerical example with the cost function in (15) in §5.3 are almost constant. In the numerical example there are other terms, but the dominating ones are the quadratic terms. As the queues get larger, the influence of the smaller-power terms decreases, and the optimal ratios converge to fixed numbers. If the function is separable (as would

be the case if our example had not had the Q_1Q_2 term), then the convergence is to the same ratios as if the only terms are $c_1Q_1^n + c_2Q_2^n$. The mixed terms of power n change these numbers. For the cost function in (15), the Q_1Q_2 terms is also of power 2, and hence the optimal ratios converge to different numbers than (EC.9). But for that example, clearly the optimal ratios are nearly constant.

In §EC.4 we introduced the general separable quadratic cost function to provide a tractable approximation for a broad range of possible cost functions. We observed that the optimal queue-ratio function becomes a shifted version of FQR-T, which is just FQR-T centered at points $k^*_{1,2}$ and $k^*_{2,1}$ instead of centered at 0. We now illustrate the resulting control for a candidate cost function. In order to make the linear components have approximately equal weight to the quadratic components when the queue lengths are about 50, we divide the coefficients c_i for the quadratic terms by 10. We also omit the mixed term Q_1Q_2, which violated the separability property. Instead of the cost function in (15), we now consider the cost function

$$C(Q_1, Q_2) \equiv 0.3Q_1^2 + 10Q_1 + 0.2Q_2^2 + 5Q_2. \tag{EC.18}$$

The centering is depicted by the y-intercepts on the two lines in Figure EC.13.

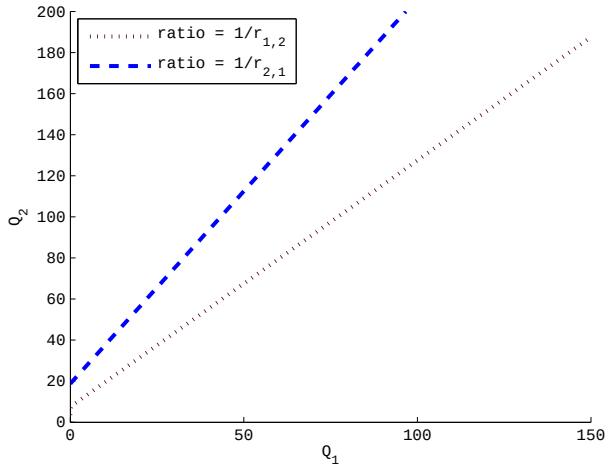

Figure EC.13 The optimal queue ratios (shifted FQR) for an X model with parameters $\mu_{i,i} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.8$, $\theta_i = 0.3$, $m_i = 100$ and the separable quadratic cost function $C(Q_1, Q_2) \equiv 0.3Q_1^2 + 10Q_1 + 0.2Q_2^2 + 5Q_2$ in (EC.18).

We now consider the two linear cases. When one or more of the component cost functions C_i is

linear, we are led to modify our control. We indicate how our fluid-model analysis can be applied to generate alternative controls in these cases, but we do not examine their performance here.

$\mathbf{n_1 = 2, n_2 = 1}$. The cost-function $C(Q_1, Q_2) = c_1 Q_1^2 + c_2 Q_2$ has one quadratic term and one linear term. The special structure of this function (C_2 not strictly convex) changes the control. Now, there is no longer dependence on the two queues, since Q_2 no longer comes into play. By (EC.16),

$$Q_1^* = \frac{a_{1,2} c_2}{2c_1} \equiv k_{1,2}^* \qquad \text{or} \qquad Q_1^* = \frac{a_{2,1} c_2}{2c_1} \equiv k_{2,1}^*.$$

Thus, we are no longer trying to keep a relation between the two queues, but instead we keep Q_1 not bigger than $k_{1,2}^*$ or $k_{2,1}^*$, depending which is optimal to use. To keep Q_1 at its optimal target, we modify our control: If class 1 is overloaded such that $q_1 > k_{1,2}^*$, then whenever $D_{1,2} \geq \max\{\kappa_{1,2}, k_{1,2}^*\}$ every newly available agent takes his next customer from the head of queue 1. Otherwise, every agent takes his next customer from the head of its own class queue.

If class 2 is overloaded, then we can have $Z_{2,1} > 0$ as long as we keep $Q_1 < k_{2,1}$. Hence, if $D_{2,1} < \kappa_{2,1}$ and $Q_1 < k_{2,1}$, then every newly available agent takes his next customer from the head of Q_2. Otherwise, he will take his next customer from the head of his own class queue.

$\mathbf{n_1 = 1, n_2 = 1}$. The purely-linear cost function $C(Q_1, Q_2) = c_1 Q_1 + c_2 Q_2$ is even more different than the functions we considered so far. However, it is well known that a linear function attains its minima on the boundaries of its domain. In our setting, this means that we either try to keep the queue that needs help at zero, or that we do not help it at all. When $Z_{1,2} > 0$, we have

$$\begin{aligned} C(Q_1, Q_2) = C(Z_{1,2}) &= c_1 \left(q_1 - \frac{s_1 \mu_{1,1}}{\theta_1} - \frac{\mu_{1,2}}{\theta_1} Z_{1,2} \right) + c_2 \left(q_2 - \frac{s_2 \mu_{2,2}}{\theta_2} + \frac{\mu_{2,2}}{\theta_2} Z_{1,2} \right) \\ &= c_1 \left(q_1 - \frac{s_1 \mu_{1,1}}{\theta_1} \right) + c_2 \left(q_2 - \frac{s_2 \mu_{2,2}}{\theta_2} \right) + \left(\frac{c_2 \mu_{2,2}}{\theta_2} - \frac{c_1 \mu_{1,2}}{\theta_1} \right) Z_{1,2}. \end{aligned}$$

Thus, if

$$\frac{c_1 \mu_{1,2}}{\theta_1} \leq \frac{c_2 \mu_{2,2}}{\theta_2} \tag{EC.19}$$

the function $C(Z_{1,2})$ is increasing, and its minima is attained when $Z_{1,2} = 0$. Otherwise, the function is decreasing, and it is optimal to take $Z_{1,2}$ as large as needed to ensure $Q_1 = 0$ (the simple calculus

gives us that $Z^*_{1,2} = m_2$, but of course class 1 may not have enough arrivals to fill both service pools). This means that we either share completely, or not share at all.

Similarly, for

$$\begin{aligned} C(Q_1, Q_2) = C(Z_{2,1}) &= c_1 \left(q_1 - \frac{s_1\mu_{1,1}}{\theta_1} - \frac{\mu_{1,1}}{\theta_1} Z_{2,1} \right) + c_2 \left(q_2 - \frac{s_2\mu_{2,2}}{\theta_2} + \frac{\mu_{2,1}}{\theta_2} Z_{2,1} \right) \\ &= c_1 \left(q_1 - \frac{s_1\mu_{1,1}}{\theta_1} \right) + c_2 \left(q_2 - \frac{s_2\mu_{2,2}}{\theta_2} \right) + \left(\frac{c_1\mu_{1,1}}{\theta_1} - \frac{c_2\mu_{2,1}}{\theta_2} \right) Z_{2,1}, \end{aligned}$$

if

$$\frac{c_2\mu_{2,1}}{\theta_2} \leq \frac{c_1\mu_{1,1}}{\theta_1} \tag{EC.20}$$

then $C(Z_{2,1})$ is increasing, and its minima is attained at $Z_{2,1} = 0$. Otherwise, $C(Z_{2,1})$ is decreasing, and it is optimal to take $Z_{2,1}$ as large as needed (and possible) to make sure that $Q_2 = 0$.

Rewriting the inefficient-sharing condition (2), we get

$$\frac{\mu_{1,1}}{\mu_{2,1}} \geq \frac{\mu_{1,2}}{\mu_{2,2}}.$$

If the two inequalities (EC.19) and (EC.20) hold together, then

$$\frac{\theta_1 c_2}{\theta_2 c_1} \leq \frac{\mu_{1,2}}{\mu_{2,2}} \quad \text{and} \quad \frac{\mu_{1,1}}{\mu_{2,1}} \leq \frac{\theta_1 c_2}{\theta_2 c_1}, \tag{EC.21}$$

but this contradicts the inefficient-sharing condition above, unless all the inequalities hold as equalities. Thus, we can have at most one of the inequalities, (EC.19) or (EC.20), hold under (2).

At first glance, it may seem from the discussion above that, when the holding cost is linear, we should not consider the system as an X model, but rather as an N model (sharing can be done in only one direction), if either (EC.19) or (EC.20) hold, or two independent I systems (no sharing at all), if none of these two inequalities hold. But that is not so. If there is spare capacity in one class, while the other class is overloaded, then it is always optimal to use the extra agents to help the overloaded class. Since we do not know what the overload incident will produce, we cannot restrict the model to an N model in advance.

Let us summarize what we have found: The cost analysis leads us to give priority to either queue 1 or queue 2. Suppose that it is optimal to give priority to queue 1. That leads us to set

the threshold for pool 2 helping class 1 at $\kappa_{1,2}=0$. In the fluid model that will either produce the desired result $Q_1=0$ or $Q_1>0$ and $Z_{1,2}=m_2$, with pool 2 devoting all its effort to class 1. There remains another case: when pool 1 has spare spare capacity. In that case, within the fluid model, if pool 2 is overloaded, then pool 1 should devote all the required spare capacity to serving class 2. We should have $Z_{2,1}=s_1\wedge q_2$. If $s_1>q_2$, then the help pool 1 provides to class 2 makes both queues empty, and there is remaining spare capacity for pool 1. On the other hand, if $s_1\le q_2$, then we have exactly $Z_{2,1}=s_1$. Overall, there are three possible end results in the fluid model: (i) $Q_1>0$ and $Z_{1,2}=m_2$, (ii) $Q_1=Q_2=0$, (iii) $Q_2>0$, $Q_1=0$ and $Z_{2,1}=s_1$.

We now must consider how to implement that control in the actual system. As indicated above, to give priority to queue 1 at all times, we can set $\kappa_{1,2}=0$, and we always allow pool 2 to help class 1, even if $Z_{2,1}>0$. The only difficulty is detecting whether or not pool 1 has spare capacity, so that we can have pool 1 helping class 2. For this purpose, we propose using a positive queue threshold for queue 2: We let an available agent in pool 1 help class 2 if, and only if, $Q_2>\kappa_{2,1}$, $Q_1=0$ and $Z_{1,2}=0$.

Since we allow pool 2 to serve class 1 all the time, we could possibly have simultaneous two-way sharing (both $Z_{1,2}>0$ and $Z_{2,1}>0$), but there should be only minimal simultaneous two-way sharing. It remains to further investigate this case.

EC.5. Additional Simulation Results

In this section we present additional simulation results.

EC.5.1. Comparing the Two Controls

We now supplement the comparison of the two controls (the fixed staffing levels versus QR-T) in §7.2 by presenting detailed simulation results. These are given in Table EC.1, including the half-width of 95% confidence intervals and a comparison of the simulation to the fluid approximation.

As stated before, for each case, we conducted 5 independent simulation runs using QR-T, and 5 independent simulation runs with a fixed $Z_{1,2}$, each run with $300,000$ arrivals. The independent replications make it possible to reliably estimate confidence intervals using the t statistic with

4 degrees of freedom. The large number of arrivals ensures that the transient behavior in the beginning of the simulation, before reaching steady state, does not affect the final simulation estimates.

		Cost (in thousands)		actual ratio			actual $Z_{1,2}$	
policy		Approx.	Sim.	Approx.	Sim.	std.	Approx.	Sim.
FQR-T	$r=1.20$	19.65	20.51 $\pm$0.64	1.20	1.07 $\pm$0.00	0.16 $\pm$0.01	15.0	15.9 $\pm$0.4
	$r=1$	19.35	20.28 $\pm$0.81	1.00	0.90 $\pm$0.00	0.13 $\pm$0.01	16.7	17.7 $\pm$0.3
	$r=0.83$	19.25	19.73 $\pm$0.64	0.83	0.76 $\pm$0.00	0.11 $\pm$0.00	18.4	18.9 0.5
	$r=0.60$	19.56	21.16 $\pm$0.77	0.60	0.56 $\pm$0.00	0.08 $\pm$0.00	21.4	22.1 $\pm$0.3
	$r=0.40$	20.75	22.31 $\pm$0.92	0.40	0.37 $\pm$0.00	0.06 $\pm$0.00	25.0	25.3 $\pm$0.3
fixed $Z_{1,2}$	$Z_{1,2}=15$	19.65	21.47 $\pm$0.57	1.20	1.52 $\pm$0.08	1.93 $\pm$0.29	15.0	15.0 $\pm$0.0
	$Z_{1,2}=17$	19.32	21.35 $\pm$0.46	0.96	1.13 $\pm$0.11	1.17 $\pm$0.45	17.0	17.0 $\pm$0.0
	$Z_{1,2}=19$	19.26	20.86 $\pm$0.37	0.78	0.87 $\pm$0.07	0.75 $\pm$0.57	19.0	19.0 $\pm$0.0
	$Z_{1,2}=22$	19.69	21.42 $\pm$0.60	0.56	0.61 $\pm$0.05	0.38 $\pm$0.04	22.0	22.0 $\pm$0.0
	$Z_{1,2}=25$	20.75	22.63 $\pm$0.86	0.40	0.42 $\pm$0.01	0.33 $\pm$0.14	25.0	25.0 $\pm$0.0

Table EC.1 Full simulation results of Figure (5)**. The 'approx' columns show the anticipated results according to the fluid model, and the 'sim.' columns show the simulation results, together with half-width confidence interval.**

We now provide additional observations about our simulation results for this example. Another important observation is that FQR-T is doing a better job in keeping the ratio between the two queues close to the desired ratio. The accuracy becomes even better when the system is larger (see the "ratio" row in Table 1 in the $n=400$ columns). We have also included a column showing the simulated standard deviations of the ratios. Note how small the standard deviations are when using FQR-T, in comparison to the standard deviations when using the fixed-$Z_{1,2}$ control. Since FQR-T is working towards keeping the ratio between the two queues fixed throughout, the ratio between the two queues at any time point is approximately $r_{1,2}$. It also makes the two queues

strongly positively correlated, which reduces the overall variance. In contrast, under the fixed-$Z_{1,2}$ control, the two queues are independent with zero correlation.

The simulated ratio was calculated as a long-run average of the ratio between the two queues throughout the simulation time. We can compare it to Q_1/Q_2 from Table 1 which appears in §7.1 (in the $n = 100$ columns) which is also approximately 0.9. These agree closely because of state-space collapse, as in Figure EC.8 discussed in §EC.2.

Finally, Table EC.1 shows that the fluid approximation tends to underestimate the actual average cost in the stochastic model. That is understandable, because the fluid model ignores stochastic fluctuations, which will tend to increase the average costs with a convex cost function. However, note that the fluid approximation does do an excellent job in describing the relative costs. In particular, the fluid model succeeds in locating the correct minima for both controls.

EC.5.2. Performance of FQR-T Under Normal Loading

In §4.1 we saw that FQR-T performs well when $\kappa_{1,2} = \kappa_{2,1} = 10$ for Example 2 with $n = 100$ servers and $\lambda_i = 99$, showing that FQR without thresholds and one-way sharing can perform poorly. To supplement those simulation results and the simulation results in §7, in this section we present additional simulation results. As in Table 1, we consider three values of n: $n = 25$, $n = 100$ and $n = 400$. We let the arrival rates in both queues be $\lambda_i^{(n)} = 0.98n$. The service-rate and abandonment-rate parameters are fixed at $\mu_{i,i} = 1.0$, $\mu_{1,2} = \mu_{2,1} = 0.8$, $\theta_i = 0.2$. We let the thresholds be $\kappa_{1,2}^{(n)} = \kappa_{2,1}^{(n)} = 0.1n$, rounded up to 3.0 for $n = 25$. We compare the results of FQR-T to the $M/M/n + M$ model, which would prevail if there were absolutely no sharing at all. As before, we see that the mean queue lengths are actually slightly smaller with FQR-T. That shows that the little sharing that takes place with FQR-T is not so bad.

Due to the symmetry of the system under the parameters we chose, there is no difference between the steady-state values of both queues and service pools. Thus, we display only Q_1 and $Z_{1,2}$. We can see that as the system becomes larger, the sharing decreases, and the queue size gets closer to the queue length in an $M/M/n + M$ model.

	n=25		n=100		n=400	
perf. meas.	I model	sim.	I model.	sim.	I model	sim.
$E[Q_1]$	5.1	4.8 $\pm$0.3	8.4	7.3 $\pm$1.0	11.3	10.5 $\pm$2.6
$E[Q_1/n]$	0.20	0.19 $\pm$0.01	0.08	0.07 $\pm$0.01	0.03	0.03 $\pm$0.01
$E[Z_{1,2}]$	–	1.3 $\pm$0.1	–	1.9 $\pm$0.2	–	1.3 $\pm$0.4
$E[Z_{1,2}/n]$	–	0.05 $\pm$0.01	–	0.02 $\pm$0.00	–	0.00 0.00

Table EC.2 A comparison of the exact I-model queues with simulation results for the steady-state performance measures of the X model in normal loading under FQR-T. The arrival rates are $\lambda_1^{(n)} = \lambda_2^{(n)} = 0.98n$ and the thresholds are $\kappa_{1,2}^{(n)} = \kappa_{2,1}^{(n)} = 0.1n$. Service rates are $\mu_{1,1} = \mu_{2,2} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.8$ and the abandonment rates are $\theta_1 = \theta_2 = 0.2$

EC.5.3. Sensitivity Analysis For the Thresholds

We now consider different values for the thresholds with and without one-way sharing. Our objective is to perform a sensitivity analysis for the thresholds for a finite system (in this case having 100 agents in each service pool), as a complimentary to the asymptotic line of reasoning in §6. The simulation results, displayed in Table EC.3, are for systems having $\lambda_i = 98$, $\mu_{1,1} = \mu_{2,2} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.8$ and $\theta_i = 0.2$, $i = 1, 2$. We vary the thresholds between $\kappa_{i,j} = 1$ and $\kappa_{i,j} = 30$, $i, j = 1, 2$. (Note that with $\kappa_{i,j} = 1$ FQR-T reduces to FQR.) For ease of exposition we take $r_{1,2} = r_{2,1} = 1$. The symmetry allows us to present the results for $E[Q_1]$ and $E[Z_{1,2}]$ only, and consider $\kappa_{1,2} = \kappa_{2,1}$. (If $r_{1,2} \neq r_{2,1}$ then the sensitivity analysis should be performed for each of the two thresholds separately.)

Table EC.3 clearly shows the benefits of using one-way sharing, since even with $\kappa_{i,j} = 1$ the performance is almost as good as when we add thresholds of size 15. However, recall that the thresholds play a vital role in our control: In addition to helping prevent unwanted sharing, they act as "overload detectors": When $D_{i,j}(t)$ first crosses the threshold $\kappa_{i,j}$, we consider class-i queue to be overloaded, and sharing is activated with pool j helping queue i.

As discussed in §6, we do not want to have the thresholds too large, as they will fail to detect small overloads. Moreover, we see that it can actually be beneficial to share a little, even when the

system is not overloaded; Observe that the average queue length in the case $\kappa_{i,j} = 30$ is larger than when $\kappa_{i,j}$ is $10, 15$ or 20. (See also the last paragraph in §4.)

Thus, in choosing the thresholds we need to make sure that under normal loadings they will not be crossed too often, but even small overloads will be detected. Here we see that any value in $\{10, \ldots, 20\}$ is reasonable, both with one-way sharing and without. To ensure that even small overloads will be detected by the thresholds, it is probably best to take $10 \leq \kappa_{i,j} \leq 15$.

Insight From the Asymptotic Analysis of the Thresholds. The asymptotic analysis in §6 helps to find good candidates for the thresholds for larger systems. In our example, we can heuristically think of 30 as being of order $O(n)$, while 10 and 15 are of a smaller order, say $O(n^{0.6})$. Then $30 = 0.3n$, $15 \approx n^{0.6}$ and $10 \approx 2/3n^{0.6}$.

This line of reasoning hints at what the thresholds should be (approximately) for a larger system having the same service and abandonment parameters. For example, if $n = 1000$ then $\kappa_{i,j} = 0.3n = 300$ is too large, but $64 \approx n^{0.6} \leq \kappa_{i,j} \leq 2/3n^{0.6} \approx 42$ are good candidates for the thresholds. The threshold values can be determined using simulations, just as in Table EC.3.

	With One-Way Sharing		Without One-Way Sharing	
perf. meas.	$E[Q_1]$	$E[Z_{1,2}]$	$E[Q_1]$	$E[Z_{1,2}]$
$\kappa_{i,j} = 1$	8.4 $\pm$0.4	2.8 $\pm$0.3	29.9 $\pm$1.7	38.2 $\pm$0.6
$\kappa_{i,j} = 5$	8.3 $\pm$0.9	2.4 $\pm$0.3	8.6 $\pm$0.6	8.1 $\pm$0.1
$\kappa_{i,j} = 10$	7.5 $\pm$0.4	1.9 $\pm$0.2	7.4 $\pm$0.6	3.6 $\pm$0.2
$\kappa_{i,j} = 15$	7.2 $\pm$0.6	1.5 $\pm$0.2	7.1 $\pm$0.6	2.1 $\pm$0.1
$\kappa_{i,j} = 20$	7.5 $\pm$0.7	1.1 $\pm$0.2	7.3 $\pm$0.7	1.3 $\pm$0.2
$\kappa_{i,j} = 30$	8.2 $\pm$0.9	0.5 $\pm$0.1	8.2 $\pm$0.7	0.5 $\pm$0.1

Table EC.3 Sensitivity analysis of the effect of the thresholds in a system with 100 agents in each pool. The arrival rates are $\lambda_1 = \lambda_2 = 98$. Service rates are $\mu_{1,1} = \mu_{2,2} = 1$, $\mu_{1,2} = \mu_{2,1} = 0.8$ and the abandonment rates are $\theta_1 = \theta_2 = 0.2$. All the results are derived from five independent simulation runs.

References

Cohen, J. W. 1982. *The Single Server Queue*, second ed., North-Holland, Amsterdam.

Eick, S. G., W. A. Massey, W. Whitt. 1993. The physics of The Mt/G/infty queue. *Oper. Res.* **41** (4) 731–742.

Gurvich, I., W. Whitt. 2007c. Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.*, forthcoming. Available at: http://www.columbia.edu/∼ww2040/recent.html

Gurvich, I., W. Whitt. 2007d. Asymptotic optimality of queue-ratio routing for many-server service systems. working paper. Available at: http://www.columbia.edu/∼ww2040/recent.html

Rockafellar, T. R. 1970. *Convex Analysis*, Princeton University Press, Princeton, N.J.

Whitt, W. 1989. Planning queueing simulations. *Management Sci.* **35** (11) 1341–1366.