OPERATIONS RESEARCH
http://dx.doi.org/10.1287/opre.1110.0960ec

e-companion

ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"Shadow-Routing Based Control of Flexible Multiserver Pools in Overload" by Alexander L. Stolyar and Tolga Tezcan, *Operations Research*, http://dx.doi.org/10.1287/opre.1110.0960.

Online Appendix

EC.1. Proof of Theorem 5.1

It suffices to show that for each server pool j, we have

$$\lim_{T\to\infty} \liminf_{r\to\infty} \frac{1}{T} \sum_i E\left[\int_0^T g_i \mu_{ij} \bar{\Psi}^r_{ij}(t)dt\right] = \sum_i \lambda_{ij} g_i = \sum_i g_i \mu_{ij} \bar{\psi}^*_{ij}, \tag{EC.1}$$

where $\psi^*_{ij} = \lambda_{ij}/\mu_{ij}$, $\sum_i \psi^*_{ij} = \beta_j$.

From this point on we will consider a single fixed pool j and drop index j from the notation; so that we write λ_i, μ_i, Y^r_i, Ψ^r_i and so on, instead of λ_{ij}, μ_{ij}, Y^r_{ij}, Ψ^r_{ij}. We denote by $A^r_i(t)$ the number of class i arrivals by time t, and by $\bar{A}^r_i(t) = A^r_i(t)/r$ its fluid-scaled version. Just to improve exposition, let us make further simplifying assumptions: we consider the version of SHADOW-RM that drops tagged customers on arrival, and assume that $\lambda_i > 0$ (recall, this means $\lambda_{ij} > 0$) for all flows $i \in \mathcal{I}$. (The proof without these simplifications is an obvious generalization.) Assume w.l.o.g. that flows with lower indices i have higher priority. Recall that,

$$\sum_{i=1}^{I} \psi^*_i = \beta, \tag{EC.2}$$

where $\psi^*_i = \frac{\lambda_i}{\mu_i}$. By Theorem 4.3 we have, w.p.1, for each i:

$$\bar{A}^r_i(t) \to \lambda_i t, \quad \text{u.o.c.}, \tag{EC.3}$$

as $r \to \infty$. Finally, since $(1/T)\int_0^T \bar{\Psi}^r_i(t)dt$ is obviously uniformly bounded by β, it will suffice to show that, for each i,

$$\frac{1}{T}\int_0^T \bar{\Psi}^r_{ij}(t)dt \to \psi^*_i \tag{EC.4}$$

in probability, as $r \to \infty$.

Let

$$\bar{X}^r(t) = \left(\bar{Y}^r(t), \bar{\Psi}^r(t)\right),$$

where $\bar{Y}^r(t) = \left(\bar{Y}^r_1(t), \ldots, \bar{Y}^r_I(t)\right)$ and $\bar{\Psi}^r(t) = \left(\bar{\Psi}^r_1(t), \ldots, \bar{\Psi}^r_I(t)\right)$.

We proceed with the proof of (EC.4). By Lemma A.1 and Proposition E.1 in Dai and Tezcan (2010) , for every subsequence of $\{r\}$, there exists further subsequence, denoted by $\{r\}$ for notational simplicity (depending on the sample path), such that w.p.1

$$\bar{X}^r(t) \to \bar{X}(t), \text{ u.o.c.,} \tag{EC.5}$$

as $r \to \infty$, for $\bar{X}(t) = (\bar{\Psi}(t), \bar{Y}(t))$, where $\bar{Y}(t) = \big(\bar{Y}_1(t), \ldots, \bar{Y}_I(t)\big)$ and $\bar{\Psi}(t) = \big(\bar{\Psi}_1(t), \ldots, \bar{\Psi}_I(t)\big)$, that satisfies the following fluid model equations

$$\bar{A}_i(t) = \bar{A}_i^q(t) + \bar{A}_i^\Psi(t) = \lambda_i t, \ \forall i \in \mathcal{I}, \tag{EC.6}$$

$$\bar{Y}_i(t) = \bar{Y}_i(0) + \bar{A}_i^q(t) - \bar{B}_i(t) - \theta_i \int_0^t \bar{Y}_i(s)ds, \ \forall i \in \mathcal{I}, \tag{EC.7}$$

$$\bar{\Psi}_i(t) = \bar{\Psi}_i(0) + \bar{A}_i^\Psi(t) + \bar{B}_i(t) - \mu_i \int_0^t \bar{\Psi}_i(s)ds, \ \forall i \in \mathcal{I}, \tag{EC.8}$$

$$\bar{I}(t) = \beta t - \sum_i \int_0^t \bar{\Psi}_i(s)ds, \tag{EC.9}$$

$$\int_0^t \bar{Y}_i(s)d\bar{I}(s) = 0, \ \forall i \in \mathcal{I}, \tag{EC.10}$$

$$\bar{Y}_i(t)\left(\beta - \sum_i \bar{\Psi}_i(t)\right) = 0, \ \forall i \in \mathcal{I}, \tag{EC.11}$$

$$\int_0^t \bar{A}_i^\Psi(s)d\left(\beta - \sum_i \bar{\Psi}_i(s)\right) = 0, \ \forall i \in \mathcal{I}, \tag{EC.12}$$

$$\bar{A}_i, \bar{A}_i^q, \bar{A}_i^\Psi, \bar{B}_i \text{ are nondecreasing, } \forall i \in \mathcal{I}, \tag{EC.13}$$

$$\bar{Y}_i(t) \geq 0, \bar{\Psi}_i(t) \geq 0, \sum_i \bar{\Psi}_i(t) \leq \beta, \ \forall i \in \mathcal{I}, \tag{EC.14}$$

and

$$\sum_{\ell=1}^i \dot{\bar{B}}_\ell(t) = \sum_{i=1}^I \mu_i \bar{\Psi}_i(t), \quad \text{if } \sum_{\ell=1}^i \bar{Y}_\ell(t) > 0, \ \forall i \in \mathcal{I}, \tag{EC.15}$$

for $t \geq 0$. By Lemma A.1 in Dai and Tezcan (2010) , $\bar{X}$, $\bar{A}_i^q$, $\bar{A}_i^\Psi$, $\bar{B}_i$ and $\bar{I}$ are Lipschitz continuous, hence differentiable a.e. (Thus (EC.15) should be understood in a.e. sense.) For the rest of the proof we only consider time points t where all these processes are differentiable, when we write expressions involving derivatives.

We next give a brief explanation of the fluid model equations (EC.6)–(EC.15), we refer to Dai and Tezcan (2010) for more details. The processes $\bar{A}_i^q(t)$ and $\bar{A}_i^\phi(t)$ can be interpreted as the number

of (or actually the amount of fluid for) class i customers who are routed to queue and service upon arrival by time t, respectively. The processes $\bar{Y}_i(t)$ and $\bar{\Psi}_i(t)$ are the number of class i customers who are waiting in queue and being served at time t, respectively. We use $\bar{B}_i(t)$ to denote the number of class i customers whose service started by time t and who had to wait in queue and $\bar{I}(t)$ to denote the total time servers have idled by time t. Equations (EC.6)–(EC.14) hold for any policy, (EC.15) on the other hand holds under the static priority policy. In words, it implies that if there are high priority customers in queue, they will be admitted to service before customers in other classes.

We note that similar to Lemma 4.1 in Atar et al. (2010) , we have

$$\bar{Y}_i(t) \le M, \tag{EC.16}$$

for all $t \ge 0$ and $i \in \mathcal{I}$, for some $M < \infty$, depending only on the values of $\bar{Y}_i(0)$ and system parameters.

We prove below that for any $\epsilon > 0$, there exists $T_\epsilon > 0$ such that

$$\bar{Y}_i(t) = 0,\ i \le I-1, \qquad \bar{Y}_I(t) \le \epsilon, \qquad \bar{\Psi}_i(t) \in (\Psi_i^* - \epsilon, \Psi_i^* + \epsilon)\,,\ \forall i, \tag{EC.17}$$

for $t \ge T_\epsilon$. Hence, for $T > T_\epsilon$ large enough, (EC.17) will imply (EC.4).

We next prove (EC.17) to complete the proof. The idea of the proof is as follows. We first focus on the highest priority class, class 1, customers. Note that for class 1, if $\bar{\Psi}_1(t) < \psi_1^*$, then $\bar{\Psi}_1(t)$ must be increasing. This follows from the fact that if there are class 1 customers waiting in the queue they will proceed to service before all the other customers in other classes by (EC.15). Using this fact we prove that $\bar{\Psi}_1(t) \ge \psi_1^* - \epsilon$, for all t large enough. Then, this is used to prove that $\bar{Y}_1(t) = 0$, for t large enough. We finally prove using these results that $\bar{\Psi}_1(t) \approx \psi_1^*$ for t large enough. Hence, for large enough t we can focus on the remaining customer classes, 2 and above, while "assuming" that server pool size available to them is $\beta - \psi_1^*$. Going in this fashion, using similar arguments to those for class 1 customers, we prove for t large enough that $\bar{Y}_i(t) = 0$ and $\bar{\Psi}_i(t) \approx \psi_i^*$, for $i = 1, \ldots, I-1$. Now, for class I, for t large enough, the remaining capacity is approximately

ψ_I^*, by (EC.2), enough to serve all arrivals to class I. From here we obtain that, for t large enough, $\bar{Y}_I(t) \approx 0$ and $\bar{\Psi}_I(t) \approx \psi_I^*$.

Next, we provide the details of the proof of (EC.17). First, we prove that, for any $\epsilon > 0$, there exists T_1' such that for $t \geq T_1'$

$$\bar{\Psi}_1(t) \geq \psi_1^* - \epsilon. \tag{EC.18}$$

If $\bar{Y}_1(t) = 0$, $\dot{\bar{Y}}_1(t) = 0$, since it attains a minimum at time t. Otherwise, if $\bar{Y}_1(t) > 0$, by (EC.15),

$$\dot{\bar{B}}_1(t) = \sum_{i=1}^{I} \mu_i \bar{\Psi}_i(t). \tag{EC.19}$$

Hence, by (EC.8), if $\bar{\Psi}_1(t) < \psi_1^* - \epsilon$,

$$\dot{\bar{\Psi}}_1(t) \geq \left(\lambda_1 \wedge \sum_{i=1}^{I} \mu_i \bar{\Psi}_i(t) \right) - \mu_1 \bar{\Psi}_1(t) > c\epsilon,$$

for some $c > 0$ independent of ϵ. Since $\bar{\Psi}_1(t) (\geq 0)$ is bounded from below, this proves (EC.18). Next we prove that there exists $T_1'' > T_1'$ such that for $t \geq T_1''$

$$\bar{Y}_1(t) = 0. \tag{EC.20}$$

Assume that $\bar{Y}_1(t) > 0$ for $t > T_1'$. Then by (EC.19)

$$\dot{\bar{Y}}_1(t) < \lambda_1 - \sum_{i=1}^{I} \mu_i \bar{\Psi}_i(t) < -c\epsilon,$$

for some (reselected) $c > 0$. Combining this with (EC.16), we have (EC.20). To complete the proof of (EC.17) for $i = 1$, we need to prove that there exists $T_1 > T_1''$ such that for $t \geq T_1$

$$\bar{\Psi}_1(t) \leq \psi_1^* + \epsilon.$$

This follows from (EC.6)–(EC.8) and the fact that $\dot{\bar{Y}}_1(t) = 0$ for $t \geq T_1''$.

Since ϵ is arbitrary, using similar arguments, we can prove (EC.17) for $i = 2, 3, \ldots, I-1$ for $t \geq T_{I-1}$, for some $T_{I-1} > 0$ large enough, in a similar fashion.

Next we handle the lowest priority class I. Fix $\epsilon > 0$ and assume that $\bar{Y}_I(t) > 0$ for some $t \geq T_{I-1}$. By (EC.6)–(EC.12),

$$\dot{\bar{Y}}_i(t) = 0 \text{ and } \dot{\bar{\Psi}}_i(t) = \lambda_i - \mu_i \bar{\Psi}_i(t), \tag{EC.21}$$

for all $i \leq I-1$ and $t \geq T_{I-1}$. Also, since $\bar{Y}_I(t) > 0$, by (EC.11), and the fact that $\bar{Y}_I(t)$ is continuous

$$\sum_{i=1}^{I} \dot{\bar{\Psi}}_i(t) = 0. \tag{EC.22}$$

This gives by (EC.17) (for $i \leq I-1$) and (EC.21) that

$$\dot{\bar{\Psi}}_I(t) \geq -c\epsilon, \tag{EC.23}$$

for some (reselected) $c > 0$ independent of ϵ. Also by (EC.11) and (EC.17)

$$\bar{\Psi}_I(t) \geq \psi_I^* - I\epsilon. \tag{EC.24}$$

Note that, by (EC.6)–(EC.8),

$$\dot{\bar{\Psi}}_I(t) + \dot{\bar{Y}}_I(t) = \lambda_I - \mu_I \bar{\Psi}_I(t) - \theta \bar{Y}_I(t). \tag{EC.25}$$

Combining this with (EC.23) and (EC.24), we have

$$\dot{\bar{Y}}_I(t) \leq c\epsilon - \theta_I \bar{Y}_I(t),$$

for some (reselected) $c > 0$ independent of ϵ. This implies with (EC.16) that there exits $T_\epsilon > T_{I-1}$ such that

$$\bar{Y}_I(t) < \epsilon \tag{EC.26}$$

for all $t \geq T_\epsilon$. The result for $\bar{\Psi}_I$ follows immediately from this and the fact that (EC.17) holds for $i \leq I-1$. The proof is complete.

EC.2. Fluid approximation for the performance of $c\mu/\theta$ policy in X-models

Although it is evident from the simulation results in Section 5 that the $c\mu/\theta$ policy is far from optimal in X-model systems, it is also possible to approximate the performance of X-model systems under this policy using a fluid model (see Perry and Whitt (2009, 2010) for a similar approach) to gain further insight why it does not work. Specifically, we provide approximations for the steady state behavior of the $c\mu/\theta$ policy in X-model systems. These approximations are based on the fluid limits of these systems and can be proven rigorously by taking the formal limits of the properly scaled stochastic processes associated with the queueing system. However, we will not attempt to prove any results, as it is beyond the scope of this paper. Yet, we note that, it has been demonstrated in the literature that fluid approximations are especially accurate for overloaded systems (Whitt (2004, 2006)).

Although it is possible to build fluid approximations for a more general parameter setting, we only focus on the case when both servers give priority to class 2 customers (recall that we use $c_i = g_i\theta_i$):

$$c_2\mu_{23}/\theta_2 > c_1\mu_{13}/\theta_1 \text{ and } c_2\mu_{24}/\theta_2 > c_1\mu_{14}/\theta_1; \tag{EC.27}$$

and the system has enough capacity to serve all class 2 customers:

$$\lambda_2 \leq \mu_{23}\beta_3 + \mu_{24}\beta_4. \tag{EC.28}$$

Let $Y_i^r(\infty)$ denote the number of class i customers in queue and $\Psi_{ij}^r(\infty)$ denote the number of class i customers receiving service from a server in pool j in steady state in the rth system. Let y_i and ψ_{ij} be the unique solution of the following equations;

$$p_{2j} = \frac{\sum_{i=1}^{2}\mu_{ij}\psi_{ij}}{\sum_{i=1}^{2}\sum_{j=3}^{4}\mu_{ij}\psi_{ij}}, \text{ for } j = 3,4 \tag{EC.29}$$

$$\sum_{i=1}^{2}\psi_{ij} = \beta_j, \text{ for } j = 3,4 \tag{EC.30}$$

$$p_{2j}\lambda_2 = \mu_{2j}\psi_{2j}, \text{ for } j = 3,4. \tag{EC.31}$$

$$y_1 = \frac{\left(\lambda_1 - \sum_{j=3}^{4} \mu_{1j}\psi_{1j}\right)}{\theta_1}, \text{ and } y_2 = 0. \tag{EC.32}$$

Under (EC.27) and (EC.28), the (random) quantities $Y_i^r(\infty)$ and $\Psi_{ij}^r(\infty)$ can be approximated by ry_i and rz_{ij}. It is possible to extract five linearly independent equations from (EC.29)-(EC.31) involving p_{21} and ψ_{ij}, $i,j = 1,2$, hence (EC.29)-(EC.32) has a unique solution.

The reasoning behind these approximations is the following; first, since the second customer class has priority over the first class and we assume that (EC.28) holds, very few customers are expected in the second queue, hence we propose the approximation $y_2 = 0$. In addition, since the system is overloaded, for r large, the probability that an incoming customer will find idle servers is close to zero. Therefore, almost all the customers will have to wait, including class 2 customers, before receiving service. Since class 2 customers have priority, a class 2 customer in queue will be routed to the first server pool that finishes service. In addition, because service times are assumed to be exponential, the probability that a server in pool j will complete service before the other pool is given by

$$\frac{\sum_{i=1}^{2} \mu_{ij}\Psi_{ij}^r(\infty)}{\sum_{i=1}^{2}\sum_{j=3}^{4} \mu_{ij}\Psi_{ij}^r(\infty)}, \quad \text{for } j = 3 \text{ and } 4,$$

Assuming $\Psi_{ij}^r(\infty)/r$ are close to being deterministic when r is large, we obtain (EC.29). Again, because the system is in steady state and all the class 2 customers are served, we have (EC.31) from the conservation of the flow of class 2 customers in and out of pool j. Equation (EC.32) again follows from conservation of the flow for class 1 customers and (EC.30) obviously must hold.

To test the quality of the proposed approximations, we compare them with estimates from simulation experiments. Using the parameters in Section 7, we obtain the following approximations by solving (EC.29)–(EC.32): $\Psi_{13}^r(\infty) \approx 12.86$, $\Psi_{23}^r(\infty) \approx 87.13$, $\Psi_{14}^r(\infty) \approx 37.13$, $\Psi_{24}^r(\infty) \approx 62.86$, $Y_1^r(\infty) \approx 161.39$, and $Y_2^r(\infty) \approx 0$. The rate customers abandon the system clearly can be approximated by $\theta_i Y_i^r(\infty)$, hence we deduce that about 64.56% of class 1 customers abandon the system in steady state. Compared to simulation results (see Table 1) this is very accurate with 1.8% error. In addition the estimates for the other quantities, obtained from simulations, are $\tilde{\Psi}_{11}^r = 13.56$,

$\tilde{\Psi}^r_{21} = 86.44$, $\tilde{\Psi}^r_{12} = 37.71$, $\tilde{\Psi}^r_{22} = 62.28$, $\tilde{Y}^r_1 = 158.91$, and $\tilde{Y}^r_2 = 5.98$, which demonstrates the accuracy of our approximations.

An interesting observation in this parameter setting is that under the $c\mu/\theta$ rule, if μ_{14} is arbitrarily small, almost all the class 1 customers abandon the system. For example if we set $\mu_{14} = 0.01$ (while keeping all the other parameters fixed), 99.5% of class 1 customers abandon the system, implying almost 15% less in revenue per unit time compared to the optimal solution of the SPP and 13.3% less than the revenue under the SHADOW-RM algorithm. When we simulate this system for 5 million arrivals, the proportion of class 1 customers abandoned from class 1 is 99.18%.

References

Atar, R., C. Giat, N. Shimkin. 2010. The $c\mu/\theta$ rule for many-server queues with abandonment. *Operations Research* **58**(5) 1427–1439.

Dai, J. G., T. Tezcan. 2010. State space collapse in many server diffusion limits of parallel server systems. *Math Oper. Res.* **36**(2) 271–320.

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

Perry, O., W. Whitt. 2010. A fluid approximation for service systems responding to unexpected overloads. Tech. rep., Columbia University.

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

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Operations Research* **54**(2) 37–54.