Probability in the Engineering and Informational Sciences, **24**, 2010, 171–200.
doi:10.1017/S0269964809990222

THE *N*-NETWORK MODEL WITH UPGRADES

DOUGLAS G. DOWN
Department of Computing and Software
McMaster University
Hamilton, ON L8S 4L7, Canada
E-mail: downd@mcmaster.ca

MARK E. LEWIS
School of Operations Research and Information Engineering
Cornell University
226 Rhodes Hall, Ithaca, NY 14853
E-mail: mark.lewis@cornell.edu

In this article we introduce a new method of mitigating the problem of long wait times for low-priority customers in a two-class queuing system. To this end, we allow class 1 customers to be *upgraded* to class 2 after they have been in queue for some time. We assume that there are c_i servers at station i, $i = 1, 2$. The servers at station 1 are flexible in the sense that they can work at either station, whereas the servers at station 2 are dedicated. Holding costs at rate h_i are accrued per customer per unit time at station i, $i = 1, 2$. This study yields several surprising results. First, we show that stability analysis requires a condition on the order of the service rates. This is unexpected since no such condition is required when the system does not have upgrades. This condition continues to play a role when control is considered. We provide structural results that include a c-μ rule when an inequality holds and a threshold policy when the inequality is reversed. A numerical study verifies that the optimal control policy significantly reduces holding costs over the policy that assigns the flexible server to station 1. At the same time, in most cases the optimal control policy reduces waiting times of *both* customer classes.

1. INTRODUCTION

In service systems (such as call centers), it may be the case that two types of customers must be served: the first, a high revenue (or high holding cost) stream and the second

 0269-9648/10 $25.00

with correspondingly lower revenues or costs. If servers are capable (trained) to serve both demand types, then the problem becomes one of assigning (static or dynamic) priorities to arriving customers. Another means to indirectly enforce priorities is to have a subset of the servers dedicated to the higher revenue arrivals. The remaining *flexible* servers are capable of working on both types of customers. This has a benefit of reducing the number of skills required by servers: The dedicated servers require only one, which means a reduction in overhead if acquiring skills comes with a cost. It also means that only the flexible servers need to be controlled. This leads itself to analysis via the already classic "N-network" (cf. [15]). In the N-network, any time the flexible servers spend on the high-priority customers is, in some sense, at the expense of the low priority customers: a quality of service concern. Thus, there remains the somewhat secondary concern of the waiting times of the lower-priority customers.

In this article we propose (and analyze) an extension of the N-network that has the potential to address the allocation question above while giving consideration to the waiting times of the lower-class customers. After having waited for some time in the low-priority queue, we *upgrade* the low-priority customers to the high-priority queue, treating them as high-priority customers thereafter. Rather than keeping track of the current waiting time of each customer in queue and upgrading based on the queue lengths, we assume that the time spent in queue before upgrades is random and exponential. The decision maker still must allocate flexible servers based on the number of customers currently in each queue, the holding costs at each station, and the arrival, service, and upgrade parameters of the model.

We view the analysis that follows to be interesting from both a managerial and theoretical standpoint. Suppose there is at most one server at each station and no upgrades. When there are no dedicated servers and several parallel stations, it is well known that a flexible server should be allocated using the c-μ rule (cf. [8]). In the two-station case, when there is a dedicated server at one queue and a flexible server at the other queue (the N-network), the c-μ rule is optimal (in the asymptotic sense) when an inequality holds [6]. There exists a monotone optimal policy when the inequality is reversed. Generalizations of these results can be found in [11,12,14,17]. All of these latter results consider heavy traffic asymptotics.

The upgrade mechanism we consider has as a goal to take advantage of idle periods of the dedicated servers, with minimal impact on the high-priority arrivals. This effect appears to be lost in a heavy traffic setting, as there is asymptotically no idling. We also note that in neither case (with or without the dedicated server) does the relationship between the service rates between the two queues appear to play an important role. We present stability conditions for the system with upgrades and show that in fact when upgrades are allowed, the relationship between the service rates plays a key role. This is somewhat surprising given the previously mentioned results and serves as a warning that if this fact is ignored, an unstable design can result.

From these results it is clear that the ordering of the service rates cannot be ignored in the control. We show that a c-μ rule holds in one direction and a monotone optimal policy exists in the other direction when there are no upgrades (even in light traffic). When upgrades are added, this result continues to hold, but only under the assumption

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

that the service rates are ordered in a specific way. We also include a numerical study that shows that the optimal policy in the system with upgrades performs well both in terms of holding costs and waiting times when compared to the policy that always prioritizes station 1. In many cases, the waiting times are improved for *both* customer classes. In summary, the main contributions of the article are as follows:

- We introduce *upgrades* as a way to balance the load between customer classes and alleviate excessive wait times of low-class customers.
- We derive stability conditions for the resulting system.
- We show that under certain conditions there exist optimal policies with structure similar to that of a model without upgrades.
- We show that the aforementioned structure of the optimal policy does not necessarily hold when the conditions are violated. In fact, stability might not even hold.
- We discuss the effect of upgrades on optimal cost and waiting times.

We do not attempt a complete review of the related literature. The interested reader should view the excellent literature surveys of Aksin, Armony, and Mehrotra [2] and Gans, Koole, and Mandelbaum [13]. In this article we consider an extension of the N-network, so we restrict attention to this model and its variants. With minor modifications, the N-network can be described via the current study by setting the upgrade rate to zero and the interarrival and service distributions to be exponential. The exponential assumption allows the formulation of the decision scenarios as a Markov decision process and significantly simplifies the fluid analysis. To our knowledge, the N-network was first introduced (from a research standpoint) by Harrison [15]. The asymptotic optimality of threshold policies in the N-network was shown by Bell and Williams [6].

Other related studies on the N-network include that of Ahn, Duenyas, and Zhang [1], in which the authors show that the optimal policy in a system without arrivals (a clearing system) exhibits either a monotone switching curve structure or is exhaustive in one of the queues. Additionally, there has been some work on *bilingual* call centers. These call centers are somewhat new to the United States, but they have been prevalent in other countries for quite some time (cf. Stanford and Grassman [19]). Although upgrades present different challenges, they are related to systems with reneging. In [20,21] Ward and Glynn showed that the queue length processes of systems with exponential interarrival service times and renege times with reneging and balking and $G/G/1$ systems with reneging and balking (under heavy traffic) can be approximated with a regulated Ornstein–Uhlenbeck process. Other possibilities for reducing congestion and dealing with reneging are to announce a delay estimate to arriving customers as described in Armony, Shimkin, and Whitt [5] or to provide callers with a call-back option as described in Armony and Maglaras [3,4]. Of course, these estimation schemes do not work when the server allocation is adjusted dynamically, as in the current model.

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

The rest of the article is organized as follows. Section 2 contains a description of the model. In order to simplify the analysis, we show that there exists an optimal nonidling policy under the finite horizon discounted cost, the infinite horizon discounted cost, and the average cost cases in Section 3. In Section 4 we provide the stability analysis. We uniformize in Section 5. This allows us to define a discrete-time Markov decision process and prove the existence of an optimal policy that follows a c-μ rule in one direction but not the other. A detailed numerical study is provided in Section 6. The article is concluded in Section 7.

2. MODEL DESCRIPTION

Consider a two-station parallel queuing system where station 1 (2) is equipped with c_1 (c_2) servers. Both stations are assumed to have infinite buffer space. Customers arrive to station 1 (2) in accordance with a Poisson process of rate λ_1 (λ_2), and service times at station 1 (2) by those servers originally assigned there are assumed to be exponential with rate μ_1 (μ_2). Servers at station 1 are flexible in the sense that they are able to work at either station. When a flexible server works at station 2, the service time is still exponential, but the rate is denoted μ_3. Furthermore, if there are more than c_1 customers at station 1, then those customers above c_1 are *upgraded* to station 2 after an exponential amount of time with parameter β. The rate of upgrades is bounded by $L\beta$, where initially L is assumed to be finite. That is to say, when the queue length at station 1 is i, customers are being upgraded at the rate $((i - c_1)^+ \wedge L)\beta$. We note that from a modeling perspective, it does not matter which customer is chosen to be upgraded; however, from an implementation standpoint, it seems that upgrading from the middle of the queue is most reasonable. Holding costs are accrued at rate $h_n > 0$ for customers at station $n, n = 1, 2$. See Figure 1.

A few notes regarding the modeling assumptions are in order. First, we have chosen to place an upper bound on the upgrade rate to provide a parameter to control the worst-case rate at which customers are being upgraded. Of course, this also simplifies the analysis. Second, again due to the fact that upgrades are a manager's choice, we have assumed that a customer that is being served (or could be served) is not upgraded. Let $\mathbb{X} = \{(i,j)|(i,j) \in \mathbb{Z}^+ \times \mathbb{Z}^+\}$ be the state space, where $\mathbb{Z}^+$ represents the nonnegative integers and i (j) represent the number of customers at station 1 (2) (including those in service). A decision maker must decide how to dynamically allocate the flexible servers based on the number of customers currently in each queue. Note that the decision maker is relieved of the burden of deciding when to upgrade customers.

There are other possibilities for the upgrade mechanism. One possibility is to make the upgrade times deterministic. Another is to enforce dynamic priorities by increasing the priorities of both arrivals based on their time in queue and using these priorities to schedule the flexible servers. Although both have their appeal from a modeling standpoint, their analysis is quite difficult. Our goal here is to show that a simple upgrade mechanism can yield significant performance gains. We would expect the other mechanisms just described to have similar (if not better) performance gains.

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

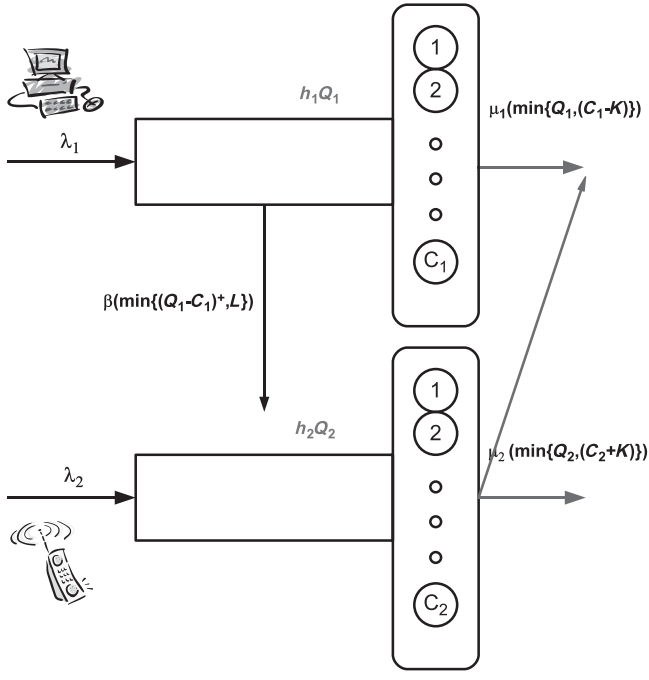

FIGURE 1. N-Network with upgrades (K is the number of servers assigned to station 2).

Consider the cost function $C((i,j),a) = ih_1 + jh_2$, where the action a denotes the number of flexible servers currently placed at station 2. We seek a policy that prescribes the action to choose for each set of queue lengths for all time. The total discounted expected cost of a policy π up to time t is

$$v_{\alpha,t}^{\pi}(i,j) = \mathbb{E}_{(i,j)} \int_0^t e^{-\alpha s} \Big(h_1 Q_1^{\pi}(s) + h_2 Q_2^{\pi}(s) \Big) ds,$$

where $Q_i^{\pi}(s)$ is the queue length (including those in service) at time s at station $i = 1, 2$ under policy π. Along with the finite horizon discounted expected cost, the other criteria considered in this article are the following:

$$v_{\alpha}^{\pi}(i,j) = \lim_{t\to\infty} v_{\alpha,t}^{\pi}(i,j) = \mathbb{E}_{(i,j)}^{\pi} \int_0^{\infty} e^{-\alpha s} \Big(h_1 Q_1^{\pi}(s) + h_2 Q_2^{\pi}(s) \Big) ds,$$

$$\rho^{\pi}(i,j) = \limsup_{t\to\infty} \frac{v_{0,t}^{\pi}(i,j)}{t},$$

where $v_{\alpha}^{\pi}(i,j)$ represents the infinite horizon α-discounted cost under π starting in (i,j) (the interchange of limit and expectation is justified by the monotone convergence

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

theorem) and $\rho^{\pi}(i,j)$ represents the long-run average expected cost per unit time. Note that in the finite horizon case, $\alpha \in [0,\infty)$, whereas in the infinite horizon case, $\alpha \in (0,\infty)$. Under any of the optimality criteria, a policy π^* is called optimal if $y^{\pi^*}(i,j) = \inf_{\pi\in\Pi} y^{\pi}(i,j)$, where Π is the set of all (measurable, nonanticipating) policies and $y = v_{\alpha,t}, v_{\alpha}$, or ρ depending on the optimality criterion.

3. NONIDLING POLICIES

In this section we show that a decision maker need not consider policies that idle flexible servers when they could be working at station 2. We begin by providing the following monotonicity result.

PROPOSITION 3.1: *Let $y = v_{\alpha}$ or $v_{\alpha,t}$ depending on the optimality criterion. For all states (i,j) the following inequalities hold:*

1. $y(i,j+1) \geq y(i,j)$,
2. $y(i+1,j) \geq y(i,j)$,

for each α, where in the finite horizon case, the result holds for all t.

PROOF: We prove the assertions in the infinite horizon case via a sample path argument. The finite horizon case is directly analogous. Consider the first result. Suppose we start two processes on the same probability space so that they see the same arrivals, (potential) services, and (potential) upgrades. Process 1 starts in state $(i,j+1)$ and uses the optimal policy, say π^*. Process 2 uses the same allocation at station 1 as Process 1 (as though it started in $(i,j+1)$) whenever possible but allows idling at station 2 if it cannot match the service rate of Process 1. For example, suppose $j+1 = c_1 + c_2$ and π^* assigns all servers to station 2. The service rate is $c_1\mu_3 + c_2\mu_2$. Since Process 1 has assigned zero servers to station 1, Process 2 does the same. However, since Process 2 has only $c_1 + c_2 - 1$ customers at station 2, it assigns only that number of servers there and idles the other server. The analysis follows as above if $j = c_2$ and $i = c_1 - 1$. If Process 1 assigns $c_1 - 1$ servers to station 1 and one to station 2, Process 2 matches the number of servers at station 1 but idles the last server. Finally, note that if $j = 0$, then Process 1 may still be serving at station 2, while Process 2 cannot. In each case, since Process 1 has more customers at station 2 than Process 2, there is the possibility that it will see an extra service (in fact eventually this will occur).

Define the event A as the event that there is a service seen by Process 1 that is not seen by Process 2. If this extra service is seen, the relative position of the two processes is now the same; they couple (using the same policy thereafter) and receive the same cost streams. Let γ be the time that event A occurs. Note that before time γ,

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

Process 1 is accruing cost at rate h_2 higher than that of Process 1. Thus,

$$\begin{aligned} v_\alpha(i,j) - v_\alpha(i,j+1) &\le v_\alpha^{\psi}(i,j) - v_\alpha(i,j+1) \\ &= -h_2 \int_0^\infty \frac{1}{\alpha}(1 - e^{-\alpha s})\, dP(\gamma \in (s, s+ds)) \le 0, \end{aligned}$$

where ψ denotes the (potentially) suboptimal policy followed by Process 2.

To prove the second result, again define two processes on the same space. This time let Process 1 start in $(i+1,j)$ and Process 2 start in (i,j). Process 2 uses the same allocation at station 2 as Process 1 (as though it started in $(i+1,j)$) whenever possible, but it allows idling at station 1 if it cannot match the service rate of Process 1. As in the proof of the previous assertion, there is the possibility that there is an extra service seen by Process 1, not seen by Process 2. There is also the possibility of an extra upgrade seen by Process 1. Let B be the event that the prior case occurs before the latter. In this case, an argument similar to the proof of the first assertion leads to the two processes coupling. On the complement of this event, Process 1 moves to a state, say $(i',j'+1)$ while Process 2 is in (i',j'). Let τ represent the time that either the processes couple or an extra upgrade is seen by Process 1. We have

$$\begin{aligned} v_\alpha(i,j) - v_\alpha(i+1,j) &\le v_\alpha^{\tilde{\psi}}(i,j) - v_\alpha(i+1,j) \\ &= -h_1\, \mathbb{E}\left[\frac{1}{\alpha}(1 - e^{-\alpha\tau})|B\right] P(B) \\ &\quad + \mathbb{E}[e^{-\alpha\tau}(v_\alpha(X_\tau, Y_\tau) - v_\alpha(X_\tau, Y_\tau + 1))|B^c]P(B^c) \le 0, \end{aligned}$$

where $\tilde{\psi}$ denotes the (potentially) sub-optimal policy used by Process 2, (X_τ, Y_τ) represents the state of the system in Process 2 immediately after time τ, and the second inequality follows using the first result of the proposition. ■

Consider a left-continuous policy f and let $d^f(w) \in \{0, 1, \ldots, c_1\}$ denote the number of the c_1 servers allocated to station 2 at time w under this policy. Note that $b^f_{1,k}(w) := 1_{\{k \le c_1 - d^f(w)\}}$ represents the indicator that the kth (flexible) server is available for service at station 1. Let $b^f_{2,k}(w) := 1_{\{k \le d^f(w)\}}$ be the indicator that the kth (flexible) server is available at station 2. Note that if we allow k to range over the natural numbers, then $b^f_{j,k}(w) = 0, j = 1,2$ for $k > c_1$. Define $A_i(s)$ to be the Poisson arrival process (rate λ_i) to queue $i, i = 1,2$. Similarly, let $Y^{\mu_i}_{k_i}$ be a Poisson process with rates μ_i for $i = 1,2,3$ and $k_1 = 1,2,\ldots,c_1$, $k_2 = 1,2,\ldots,c_2$, and $k_3 = 1,2,\ldots,c_1$. These represent the (potential) services of each server at each station. Denote by $Y^{\beta}_{k_3}$ for $k_3 = 1,2,\ldots,L$ the Poisson processes of rate β that model the (potential) upgrades

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

from queue 1 to queue 2. The queue length processes can be written

$$Q_1^f(s) = Q_1^f(0) + A_1(s) - \sum_{k_1=1}^{c_1} \int_0^s (1_{\{Q_1^f(w-)\geq k_1\}} b_{1,k_1}^f(w))\, dY_{k_1}^{\mu_1}(w)$$
$$- \sum_{k_3=1}^{L} \int_0^s \left(1_{\{Q_1^f(w-)\geq c_1+k_3\}}\right) dY_{k_3}^{\beta}(w)$$

and

$$Q_2^f(s) = Q_2^f(0) + A_2(s) - \sum_{k_2=1}^{c_2} \int_0^s \left(1_{\{Q_2^f(w-)\geq k_2\}}\right) dY_{k_2}^{\mu_2}(w)$$
$$- \sum_{k_2=1}^{c_1} \int_0^s \left(1_{\{Q_2^f(w-)\geq c_2+k_2\}} b_{2,k_2}^f(w)\right) dY_{k_2}^{\mu_3}(w)$$
$$+ \sum_{k_3=1}^{L} \int_0^s \left(1_{\{Q_1^f(w-)\geq c_1+k_3\}}\right) dY_{k_3}^{\beta}(w).$$

Consider

$$h_1\, \mathbb{E}\, Q_1^f(s) = h_1\, \mathbb{E}\, Q_1^f(0) + \lambda_1 h_1 s - h_1\mu_1 \sum_{k_1=1}^{c_1} \int_0^s \mathbb{E}\left(1_{\{Q_1^f(w-)\geq k_1\}} b_{1,k_1}^f(w)\right) dw$$
$$- h_1\beta \sum_{k_3=1}^{L} \int_0^s \mathbb{E}\left(1_{\{Q_1^f(w-)\geq c_1+k_3\}}\right) dw$$

and

$$h_2\, \mathbb{E}\, Q_2^f(s) = h_2\, \mathbb{E}\, Q_2^f(0) + \lambda_2 h_2 s - h_2\mu_2 \sum_{k_2=1}^{c_2} \int_0^s \left(\mathbb{E}\, 1_{\{Q_2^f(w-)\geq k_2\}}\right) dw$$
$$- h_2\mu_3 \sum_{k_2=1}^{c_1} \int_0^s \mathbb{E}\left(1_{\{Q_2^f(w-)\geq c_2+k_2\}} b_{2,k_2}^f(w)\right) dw$$
$$+ h_2\beta \sum_{k_3=1}^{L} \int_0^s \left(1_{\{Q_1^f(w-)\geq c_1+k_3\}}\right) dw,$$

where the fact that we can replace the Poisson processes with their rates follows, by, for example, Brémaud [7] (see the "Partial result" on page 24 of [7]). Assuming

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

$L \geq \max\{c_1, c_2\}$, let

$$\phi^f(s) = \sum_{k=1}^{L} \int_0^s \Big(\beta(h_2 - h_1)\, \mathbb{E}\, 1_{\{Q_1^f(w-) \geq c_1 + k\}} - \mu_1 h_1\, \mathbb{E}\, 1_{\{Q_1^f(w-) \geq k\}} b_{1,k}^f(w) - \mu_3 h_2\, \mathbb{E}\left(1_{\{Q_2^f(w-) \geq c_2 + k\}} b_{2,k}^f(w)\right)\Big)\, dw - \mu_2 h_2 \sum_{k=1}^{c_2} \int_0^s \left(\mathbb{E}\, 1_{\{Q_2^f(w-) \geq k\}}\right) dw. \tag{3.1}$$

We thus have the following alternative form for the total expected discounted cost:

$$v_{\alpha,t}^{\pi}(i,j) = (ih_1 + jh_2)\left(\frac{1}{\alpha}\right)(1 - e^{-\alpha t}) + (\lambda_1 h_1 + \lambda_2 h_2)\left[\left(\frac{1}{\alpha}\right)^2 - \left(\frac{e^{-\alpha t}}{\alpha}\right)\left(t + \frac{1}{\alpha}\right)\right] + \int_0^t e^{-\alpha s}\phi^{\pi}(s)\, ds.$$

This leads to the first major result that will considerably simplify the Markov decision process formulation to follow.

PROPOSITION 3.2: *There exists an optimal policy that does not allow for unforced idling.*

PROOF: The result is proved via a sample path argument. Suppose the system is initially in state (i,j). Consider an arbitrary policy π and a fixed time w. Suppose $Q_1^{\pi}(w-) < c_1$ and $Q_2^{\pi}(w-) > c_2$. Thus, immediately prior to w (at time $w-$), there are more customers at station 2 than can be handled by the c_2 dedicated servers, but not enough work for all servers to be busy at station 1. This implies that the first term of the integrand in (3.1) is zero. Similarly, for any k such that $Q_1^{\pi}(w-) < k \leq c_1$ corresponding to a potentially idle server, the second term of the integrand is also zero. Suppose π idles server k while another policy, say π', follows precisely the same actions of π, but at time $w-$, it allows the kth server to work until either one customer service is completed or π ceases to idle server k (whichever occurs first). Let the time of this event be $w + S_1$. There are several cases to consider. If $w + S_1 > t$ or $w + S_1$ does not represent a time of a service completion by the kth server, then the queue length processes under π and π' remain the same on $(w, w + S_1)$. Since the third term of (3.1) is nonpositive, an optimal policy is minimized when $b_{2,k}^{\pi}(w)$ is maximized. That is, when extra servers are assigned to station 2. After this time, if the policies coincide, then by (3.1) we have $v_{\alpha,t}^{\pi}(i,j) \geq v_{\alpha,t}^{\pi'}(i,j)$. If, on the other hand, $w + S_1$ corresponds to a service completion by the kth server, then the process under π' moves to some state $(i', j' - 1)$ while the process under π remains in state (i', j'). Assuming the processes follow the same policy from then on yields $v_{\alpha,t}^{\pi}(i,j) \geq v_{\alpha,t}^{\pi'}(i,j)$ since $v_{\alpha,t-(w+S_1)}^{\pi}(i',j') \geq v_{\alpha,t-(w+S_1)}^{\pi}(i',j'-1)$ (almost surely) by Proposition 3.1. Thus, a policy that does not allow unforced idling is optimal as desired. ■

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

We conclude this section by noting that the results of Propositions 3.1 and 3.2 do not require that $L < \infty$.

4. STABILITY

In this section we consider the stability of the system. In particular, we are interested in determining conditions under which the system admits a policy that has finite average cost (which implies existence of a stationary distribution). Let $\{S_{i,j}(t), t \geq 0\}$ denote the Poisson processes consisting of independent and identically distributed (i.i.d.) processing times for servers of type j at queue i, for $(i,j) \in \{(1,1),(2,1),(2,2)\}$. Let $U(t)$ be the number of customers that have been upgraded in $[0,t]$ and suppose $T_{i,j}(t)$ is the total time all servers of type j have worked at queue i in $[0,t]$. For example, $T_{1,1}(t)$ is bounded by $c_1 t$. We can now rewrite the queue length processes (suppressing the dependence on the policy f) as

$$Q_1(t) = A_1(t) - U(t) - S_{1,1}(T_{1,1}(t)), \tag{4.1}$$

$$Q_2(t) = A_2(t) + U(t) - S_{2,2}(T_{2,2}(t)) - S_{2,1}(T_{2,1}(t)). \tag{4.2}$$

Of course, $T_{i,j}(t)$ depends on the policy chosen. Rather than providing $T_{i,j}(t)$ in complete detail, we describe them sufficiently to provide stability conditions. To do this, we will use the standard technique of fluid model analysis (see Dai [9]). Suppose $\{x_n, n \geq 0\}$ is a sequence of nonnegative real numbers such that $\lim_{n\to\infty} x_n = \infty$. Define a fluid limit $(\bar{Q}(t), \bar{T}(t))$, where $\bar{Q}(t) = (\bar{Q}_1(t), \bar{Q}_2(t))$, $\bar{T}(t) = (\bar{T}_{1,1}(t), \bar{T}_{2,2}(t), \bar{T}_{2,1}(t))$, and

$$\bar{Q}_k(t) = \lim_{n\to\infty} x_n^{-1} Q_k(x_n t),$$

$$\bar{T}_{i,j}(t) = \lim_{n\to\infty} x_n^{-1} T_{i,j}(x_n t).$$

From Theorem 4.1 of [10], we have that

$$\bar{A}_i(t) = \lim_{n\to\infty} x_n^{-1} A_i(x_n t) = \lambda_i t,$$

$$\lim_{n\to\infty} x_n^{-1} S_{i,j}(T_{i,j}(x_n t)) = \mu_i \bar{T}_{i,j}(t).$$

Scaling (4.1) and (4.2) yields

$$\bar{Q}_1(t) = \lambda_1 t - \bar{U}(t) - \mu_1 \bar{T}_{1,1}(t), \tag{4.3}$$

$$\bar{Q}_2(t) = \lambda_2 t + \bar{U}(t) - \mu_2 \bar{T}_{2,2}(t) - \mu_3 \bar{T}_{2,1}(t), \tag{4.4}$$

where $\bar{U}(t) = \lim_{n\to\infty} x_n^{-1} U(x_n t)$.

We first examine the stability of queue 1. We will later connect this result to the stability of the entire system for the underlying queuing network model. The arrival

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

rate is λ_1 and the maximum departure rate is $c_1\mu_1 + L\beta$, due to the combined effects of service completions and upgrades. Intuitively, one would expect that if the arrival rate is less than the maximum departure rate, then queue 1 can be stabilized. The following lemma makes this precise in terms of fluid limits. We postpone the connection to the underlying queuing model.

PROPOSITION 4.1: *Consider the following cases:*

1. *Suppose* $\lambda_1 < c_1\mu_1 + L\beta$. *If either*
   (a) $\lambda_1 < L\beta$ *or*
   (b) $\lambda_1 \geq L\beta$
   and for some $\varepsilon > 0$, $d(\bar{T}_{1,1}(t))/dt > (\lambda_1 - L\beta + \varepsilon)/\mu_1$ *whenever* $\bar{Q}_1(t) > 0$, *then there exists* $t_1 < \infty$ *such that* $\bar{Q}_1(t) = 0$ *for all* $t \geq t_1$.
2. *If* $\lambda_1 > c_1\mu_1 + L\beta$, *then* $\bar{Q}_1(t) \to \infty$ *as* $t \to \infty$.

PROOF: Suppose $\bar{Q}_1(t) > 0$. As $U(t)$ is bounded above by a renewal process of rate $L\beta$, we have $\bar{U}(t) - \bar{U}(s) \leq L\beta(t-s)$. Combining this with the fact that $\bar{T}_{1,1}(t) - \bar{T}_{1,1}(s) \leq c_1(t-s)$ yields that $\bar{Q}_1(t)$ is continuous. So, there exists an $h > 0$ such that $\min_{s\in[t,t+h]} \bar{Q}_1(s) := c > 0$. Now, from Theroem 4.1 of [10], there exists a subsequence $\{x_{n_k}, n_k \geq 0\}$ such that $\{x_{n_k}^{-1} Q_1(x_{n_k}s), n_k \geq 0\}$ converges uniformly to $\bar{Q}_1(s)$ for s in $[t, t+h]$. This implies that for large enough n_k, we have $x_{n_k}^{-1} Q_1(x_{n_k}s) \geq c/2$; thus, the same $\{x_{n_k}\}$ can be chosen such that $Q_1(u) > L$ for $u \in [x_{n_k}t, x_{n_k}(t+h)]$. As $Q_1(u)$ remains above L in a neighborhood of $x_{n_k}t$, we have that $d(\bar{U}(t))/dt = L\beta$ and, thus, from (4.3),

$$\frac{d}{dt}\bar{Q}_1(t) = \lambda_1 - L\beta - \mu_1 \frac{d}{dt}\bar{T}_{1,1}(t).$$

Under the conditions in statement 1, we see that we have $d(\bar{Q}_1(t))/dt < 0$ and the result follows with $t_1 \leq \bar{Q}_1(0)/(\lambda_1 - L\beta)$ under part (a) (as $d(\bar{T}_{1,1}(t))/dt \geq 0$) and $t_1 \leq \bar{Q}_1(0)/\varepsilon$ under part (b). Statement 2 follows upon noting that $d(\bar{T}_{1,1}(t))/dt \leq c_1$ and, thus, if $\lambda_1 > c_1\mu_1 + L\beta$, $d(\bar{Q}_1(t))/dt > 0$ for all t. ■

We now turn our attention to the entire system. As we have already examined for station 1, it remains to consider station 2 under the assumption of stability of station 1. However, system stability of station 2 is a little more subtle than that of station 1.

If $h_1 = h_2$, it seems intuitive that if $\mu_3 \geq (\leq)\mu_1$, the flexible servers should prefer to serve customers at station 2 (1) since it can serve them faster there. A policy, say π^s, satisfying the following properties exhibits preference for serving at station 2 while maintaining the stability of station 1:

1. If $\bar{Q}_1(t) > 0$, all flexible servers work at station 1.
2. If $\bar{Q}_1(t) = 0$, each flexible server works at station 1 a proportion of time equal to $\max((\lambda_1 - L\beta)/(c_1\mu_1), 0)$.

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

PROPOSITION 4.2: *Suppose π^s is used. Assume $\lambda_1 < c_1\mu_1 + L\beta$.*

1. *If $\lambda_1 < L\beta$ and $\lambda_1 + \lambda_2 < c_1\mu_3 + c_2\mu_2$, then $\bar{Q}_2(t) \to 0$ as $t \to \infty$.*
2. *If $\lambda_1 \geq L\beta$ and $\lambda_2 + L\beta < c_1\mu_3 + c_2\mu_2 - (\mu_3/\mu_1)(\lambda_1 - L\beta)$, then $\bar{Q}_2(t) \to 0$ as $t \to \infty$.*

PROOF: To prove statement 1, note that since $\lambda_1 < L\beta$, π^s calls for the flexible servers to not work at station 1 if $\bar{Q}_1(t) = 0$ (work at station 1 is handled by the upgrades on the fluid scale). Moreover, by Proposition 4.1, we have for some t_1 that $\bar{Q}_1(t) = 0$ for all $t \geq t_1$. Combining these two facts, from (4.3), $d(\bar{U}(t))/dt = \lambda_1$ for $t \geq t_1$. Assume $\bar{Q}_2(t') > 0$ for some $t' \geq t_1$. Since the flexible servers work at station 2 we have $d(\bar{T}_{2,1}(s)|_{s=t'})/ds = c_1$ and $d(\bar{T}_{2,2}(s)|_{s=t'})/ds = c_2$. Substituting into (4.4), $d(\bar{Q}_2(s)|_{s=t'})/ds = \lambda_2 + \lambda_1 - c_1\mu_3 - c_2\mu_2$, and the result follows immediately.

Consider now statement 2. For π^s, if $\bar{Q}_1(t) > 0$, $d(\bar{T}_{1,1}(t))/dt = c_1$. Thus, since $\lambda_1 < c_1\mu_1 + L\beta$ by Proposition 4.1, part 1(b), we have for some t_1 and all $t \geq t_1$, $\bar{Q}_1(t) = 0$. Now, for $t \geq t_1$, from (4.3) and the second property of π^s, $d(\bar{U}(t))/dt = L\beta$. Assume $\bar{Q}_2(t') > 0$ for some $t' \geq t_1$. We have $d(\bar{T}_{2,2}(s)|_{s=t'})/ds = c_2$ and $d(\bar{T}_{2,1}(s)|_{s=t'})/ds = c_1(1 - (\lambda_1 - L\beta)/(c_1\mu_1))$. Substituting in (4.4),

$$\frac{d}{ds}\bar{Q}_2(s)|_{s=t'} = \lambda_2 + L\beta - c_2\mu_2 - c_1\mu_3\left(1 - \frac{\lambda_1 - L\beta}{c_1\mu_1}\right)$$

and the result follows immediately. ■

Consider the following sets of conditions.

Conditions (A1): $\lambda_1 < c_1\mu_1 + L\beta$ and either

1. $\lambda_1 < L\beta$ and $\lambda_1 + \lambda_2 < c_1\mu_3 + c_2\mu_2$

or

2. $\lambda_1 \geq L\beta$ and $\lambda_2 + L\beta < c_1\mu_3 + c_2\mu_2 - (\mu_3/\mu_1)(\lambda_1 - L\beta)$.

Conditions (B1): Any of the following hold:

1. $\lambda_1 > c_1\mu_1 + L\beta$,

or

2. $\lambda_1 < L\beta$ and $\lambda_1 + \lambda_2 > c_1\mu_3 + c_2\mu_2$,

or

3. $\lambda_1 \geq L\beta$ and $\lambda_2 + L\beta > c_1\mu_3 + c_2\mu_2 - (\mu_3/\mu_1)(\lambda_1 - L\beta)$.

Note that conditions (A1) and (B1) are not exactly complementary (they are missing the *equals* cases). The next theorem provides necessary and sufficient conditions

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

(with the caveat of the missing *equals* cases) for the cost function to be finite and is immediate from the previous results.

THEOREM 4.3: *The following hold:*

1. *If conditions* (A1) *hold, then there exists a server assignment policy* π *such that* $\rho^{\pi}(i,j) = \rho < \infty$ *for all initial states* (i,j).
2. *Assume* $\mu_3 \geq \mu_1$. *If conditions* (B1) *hold, then for any policy* π, $\rho^{\pi}(i,j) = \infty$ *for all initial states* (i,j).

PROOF: To prove statement 1, let each flexible server use π^s. From the proof of Proposition 4.2, we have $\bar{Q}_1(t), \bar{Q}_2(t) \to 0$ as $t \to \infty$. The result then follows from Theorem 4.1 of [10].

To show the second statement, consider the workload function

$$\bar{W}(t) = \frac{\bar{Q}_1(t)}{\mu_1} + \frac{\bar{Q}_2(t)}{\min\{\mu_2, \mu_3\}}.$$

We show that in all three cases of conditions (B1), $\bar{W}(t) \to \infty$ and apply Theorem 2.5.1 of [9]. If statement 1 of conditions (B1) hold, by Lemma 4.1, we have $\bar{Q}_1(t) \to \infty$ and, thus, $\bar{W}(t) \to \infty$. If statement 2 of conditions (B1) hold, we consider two cases. First, if $\mu_3 \leq \mu_2$,

$$\begin{aligned}
\bar{W}(t) &= \left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_3}\right) t + \bar{U}(t)\left(\frac{1}{\mu_3} - \frac{1}{\mu_1}\right) - (\bar{T}_{1,1}(t) + \frac{\mu_2}{\mu_3}\bar{T}_{2,2}(t) + \bar{T}_{2,1}(t)) \\
&\geq \left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_3}\right) t + \lambda_1 t \left(\frac{1}{\mu_3} - \frac{1}{\mu_1}\right) - c_1 t - \frac{\mu_2}{\mu_3} c_2 t \\
&\geq \left(\frac{\lambda_1}{\mu_3} + \frac{\lambda_2}{\mu_3}\right) t - c_1 t - \frac{\mu_2}{\mu_3} c_2 t,
\end{aligned}$$

where the second inequality follows from the facts that $\bar{U}(t) \leq \lambda_1 t$ (recall $\mu_3 \geq \mu_1$), $\bar{T}_{1,1}(t) + \bar{T}_{2,1}(t) \leq c_1 t$, and $\bar{T}_{2,2}(t) \leq c_2 t$. If $\mu_2 < \mu_3$, the result follows in a similar manner. We thus conclude that $\bar{W}(t) \to \infty$. Under statement 3 of conditions (B1), we repeat the steps for statement 2 of conditions (B1), where now $\bar{U}(t)$ is bounded above by $L\beta t$. For all three statements in conditions (B1), we have $\bar{W}(t) \to \infty$ and the result follows from Theorem 2.5.1 of [9]. ■

In the case that $\mu_3 \geq \mu_1$, Theorem 4.3 provides necessary and sufficient conditions for stability. It remains to consider the case when $\mu_1 > \mu_3$. We cannot expect that π^s has finite average cost since it prioritizes station 2. As when $\mu_3 \geq \mu_1$, we first analyze station 1 if the flexible server gives priority to customers there.

Under the policy that serves at station 1 except to avoid idling, the number of customers at station 1 (including the one in service) is a birth–death process with birth rate λ_1 everywhere and death rate $n\mu_1$ when there are $n \leq c_1$ customers in

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

the system, $c_1\mu_1 + (n - c_1)\beta$ when there are $c_1 + 1 \le n \le L + c_1$ customers in the system, and $c_1\mu_1 + L\beta$ when $n \ge L + c_1 + 1$. If $\lambda_1 < c_1\mu_1 + L\beta$, then the steady-state probability that there are n customers at station 1, p_n, exists. In the usual manner we have

$$p_n = \begin{cases} \dfrac{\lambda_1^n}{n!\mu_1^n} p_0, & 1 \le n \le c_1 \\ \dfrac{\lambda_1^n}{c_1!\mu_1^{c_1} \prod_{i=1}^{n-c_1} (\mu_1 + i\beta)} p_0, & c_1 + 1 \le n \le L + c_1 \\ \dfrac{\lambda_1^n}{c_1!\mu_1^{c_1} (c_1\mu_1 + L\beta)^{n-L-c_1-1} \prod_{i=1}^{L} (c_1\mu_1 + i\beta)} p_0, & n \ge L + c_1 + 1. \end{cases}$$

Solving for p_0,

$$p_0 = \left(1 + \sum_{n=1}^{c_1} \frac{\lambda_1^n}{n!\mu_1^n} + \sum_{n=c_1+1}^{L+c_1} \frac{\lambda_1^n}{c_1!\mu_1^{c_1} \prod_{k=1}^{n-c_1} (\mu_1 + k\beta)} + \frac{\lambda_1^{L+c_1+1}}{(c_1\mu_1 + L\beta - \lambda_1) c_1!\mu_1^{c_1} \prod_{\ell=1}^{L} (c_1\mu_1 + \ell\beta)} \right)^{-1}.$$

Consider the following sets of conditions:

Conditions (*A*2): $\lambda_1 < c_1\mu_1 + L\beta$ and

$$\lambda_2 + \sum_{n=c_1+1}^{L+c_1} (n-1)\beta p_n + L\beta p_0 \frac{\lambda_1^{L+c_1+1}}{(c_1\mu_1 + L\beta - \lambda_1) c_1!\mu_1^{c_1} \prod_{\ell=1}^{L} (c_1\mu_1 + \ell\beta)} < c_2\mu_2 + p_0 c_1\mu_3. \quad \textbf{(4.5)}$$

Conditions (*B*2): Either

1. $\lambda_1 > c_1\mu_1 + L\beta$

or

2. $\lambda_1 < c_1\mu_1 + L\beta$ and

$$\lambda_2 + \sum_{n=c_1+1}^{L+c_1} (n-1)\beta p_n + L\beta p_0 \frac{\lambda_1^{L+c_1+1}}{(c_1\mu_1 + L\beta - \lambda_1) c_1!\mu_1^{c_1} \prod_{\ell=1}^{L} (c_1\mu_1 + \ell\beta)} > c_2\mu_2 + p_0 c_1\mu_3.$$

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

THEOREM 4.4: *The following hold:*

1. *If conditions* (*A*2) *hold, then there exists a server assignment policy* π *such that* $\rho^{\pi}(i,j) = \rho < \infty$ *for all initial states* (i,j).
2. *Assume* $\mu_1 > \mu_3$. *If conditions* (*B*2) *hold, then for any policy* π, $\rho^{\pi}(i,j) = \infty$ *for all initial states* (i,j).

PROOF: Let the flexible servers give priority to station 1 and consider statement 1. By Proposition 4.1, $\lambda_1 < \mu_1 + L\beta$ implies that there exists t_1 such that $\bar{Q}_1(t) = 0$ for all $t \geq t_1$. Using a renewal-reward argument, we have for $t \geq t_1$,

$$\bar{U}(t) - \bar{U}(t_1) = (t - t_1)\left(\sum_{n=2}^{L+1}(n-1)\beta p_n + L\beta p_0 \frac{\lambda_1^{L+c_1+1}}{(c_1\mu_1 + L\beta - \lambda_1)c_1!\mu_1^{c_1}\prod_{\ell=1}^{L}(c_1\mu_1 + \ell\beta)}\right)$$

and

$$\bar{T}_{1,1}(t) - \bar{T}_{1,1}(t_1) = c_1(t - t_1)(1 - p_0).$$

Now, if $\bar{Q}_2(t) > 0$, we see that

$$\frac{d}{dt}\bar{Q}_2(t) = \lambda_2 + \sum_{n=2}^{L+1}(n-1)\beta p_n + L\beta p_0 \frac{\lambda_1^{L+c_1+1}}{(c_1\mu_1 + L\beta - \lambda_1)c_1!\mu_1^{c_1}\prod_{\ell=1}^{L}(c_1\mu_1 + \ell\beta)} - c_2\mu_2 - c_1 p_0 \mu_3,$$

and the result follows directly from Theorem 2.5.1 of [9].

To see statement 2, the policy that has the flexible servers give priority to station 1 trivially minimizes $Q_1(t)/\mu_1 + Q_2(t)/\mu_3$ (it minimizes the combined idle time of the servers that can only work at queue 2). Combined with the proof of statement 1, this then immediately yields statement 2. ■

5. THE STRUCTURE OF OPTIMAL POLICES

In this section we discuss optimal policy structure. Since we have assumed that $L < \infty$, we can apply *uniformization* with uniformization constant $\Psi = 1 \geq \lambda_1 + \lambda_2 + (c_1 + c_2)\max\{\mu_1, \mu_2, \mu_3\} + L\beta$ in the spirit of Lippman [16] so that instead of considering the continuous-time problems described earlier, we consider the discrete-time equivalents; that is, optimal policies remain unchanged while values coincide up to a multiplicative constant. Since we have uniformized, we restrict attention to the infinite horizon discounted cost and the average cost cases. We model the decision scenario as

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

a Markov decision process. For a policy $\pi = \{d_1, d_2, \ldots\}$ and for initial queue lengths $x = (i,j)$, define

$$v_{N,\theta}^{\pi}(x) := \mathbb{E}_x^{\pi} \sum_{n=0}^{N-1} [\theta^n C(X_n, d_n(X_n))], \tag{5.1}$$

$$v_{\theta}^{\pi} := \lim_{N\to\infty} v_{N,\theta}^{\pi}(x), \tag{5.2}$$

$$w^{\pi}(x) := \limsup_{N\to\infty} \frac{1}{N} v_{N,1}^{\pi}(x), \tag{5.3}$$

where $\{X_n, n \geq 0\}$ denotes the stochastic process representing the queue lengths at decision epoch n. Equations (5.1), (5.2), and (5.3) define the N-stage expected discounted cost, the infinite horizon expected discounted cost, and the long-run average expected cost, respectively. In the finite horizon problem, only the portion of the policy required for the time horizon is used. Moreover, this has the interpretation of considering an infinite horizon problem with only a finite number of decision epochs possible. Again, in each case we define the optimal values $y(i,j) := \inf_{\pi\in\Pi} y^{\pi}(i,j)$, where $y = v_{N,\theta}, v_{\theta}$, or w depending on the optimality criterion.

One might note that, as described, we have defined policies to only include those that are Markovian. However, our assumptions imply that this set is actually sufficient for finding an optimal policy among all nonanticipating policies in the sense that when searching over either set the optimal values coincide.

5.1. The Multi Server Case

To ease notation, let $g(i) := ((i - c_1)^+ \wedge L)\beta$. It is well known that for each $n \geq 0$, the value $v_{n,\theta}$ satisfies the *finite horizon optimality equations* (FHOEs):

$$\begin{aligned} v_{n+1}(i,j) &= ih_1 + jh_2 + \theta\Big(\lambda_1 v_n(i+1,j) + \lambda_2 v_n(i,j+1) \\ &\quad + g(i)v_n(i-1,j+1) + \min\{j, c_2\}\mu_2 v_n(i,j-1) \\ &\quad + \min_{k\in\{0,1,\ldots,c_1\}} \Big\{ \min\{i, c_1 - k\}\mu_1 v_n(i-1,j) + k1_{\{j-c_2 k\}}\mu_3 v_n \\ &\qquad \times (i,j-1) + [1 - (\lambda_1 + \lambda_2 + g(i) \\ &\qquad + \min\{i, c_1 - k\}\mu_1 + \min\{j, c_2\}\mu_2 \\ &\qquad + k1_{\{j-c_2\geq k\}}\mu_3)]v_n(i,j) \Big\}\Big). \end{aligned} \tag{5.4}$$

Similarly, if we replace v_n and v_{n+1} with v, we have the *discounted cost optimality equations* (DCOEs) and the optimal value $v_{\theta} = v$. In either case, a policy that achieves

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

the minimum is optimal under the respective criterion. We note that analogous results hold in the average cost case, but the stability of the system must be considered. Consider the *average cost optimality equations* (ACOEs):

$$w + u(i,j) \tag{5.5}$$
$$= ih_1 + jh_2 + \lambda_1 u(i+1,j) + \lambda_2 u(i,j+1) + g(i)u(i-1,j+1)$$
$$+ \min\{j, c_2 + k\}\mu_2 u(i,j-1)$$
$$+ \min_{k\in\{0,1,\ldots,c_1\}} \{\min\{i, c_1 - k\}\mu_1 u(i-1,j) + k1_{\{j-c_2\geq k\}}\mu_3 u(i,j-1)$$
$$+ [1 - (\lambda_1 + \lambda_2 + g(i) + \min\{i, c_1 - k\}\mu_1$$
$$+ \min\{j, c_2\}\mu_2 + k1_{\{j-c_2\geq k\}}\mu_3)]u(i,j)\}. \tag{5.6}$$

It should be clear that under any nonidling policy, all states communicate. It is well known (at least when all states communicate) that if a solution (w,u) to the ACOEs exists, w is the optimal average cost (independent of the initial state x) and u, called a relative value function, is unique up to an additive constant. The next proposition uses the results of Section 4 to provide conditions for convergence of the values and policies in the infinite horizon discounted cost case to those in the average case.

PROPOSITION 5.1: *Suppose either conditions* (A1) *or* (A2) *hold. The following hold:*

1. *The optimal average cost can be computed by* $\bar{w} = \lim_{\theta\uparrow 1}(1-\theta)v_\theta(i,j)$ *for any* $(i,j) \in \mathbb{X}$.
2. *There exists a subsequence* $\{\theta_n, n \geq 0\}$ *such that* $\theta_n \uparrow 1$ *and* $u_{\theta_n}(i,j) := v_{\theta_n}(i,j) - v_{\theta_n}(0,0) \to u(i,j)$, *where u is a relative value function such that* $(\bar{w}, u)$ *satisfy the ACOEs.*

PROOF: Note that for any $U > 0$, the set $\{(i,j)|ih_1 + jh_2 \leq U\}$ is finite. Moreover, under the hypotheses of the proposition, we have the existence of a policy with constant finite average cost. This implies that the assumptions in Corollary 7.5.10 of [18] hold. The results now follow from Theorems 7.2.3 and 7.5.6 of [18]. ■

The next result states that all of the flexible servers should be allocated to the station that is of higher priority while staying within the class of nonidling policies.

THEOREM 5.2: *In the finite or infinite horizon discounted cost cases and in the average cost case under the hypotheses of Proposition* 5.1, *there exists an optimal policy that satisfies the following properties:*

1. *Servers are not allocated to station* 2 *unless there are excess customers in the buffer of station* 2 ($j > c_2$).
2. *If there is enough excess capacity at station* 1 *to serve all of the customers at station* 2 ($j - c_2 \leq (c_1 - i)^+$), *assign all excess capacity to station* 2. ($k = c_1 - i$).

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

3. *If there are more customers at station* 2 *than can be served with the excess capacity at station* 1 ($(j-c_2)^+ \geq c_1 - i$), *there are but two possibilities:*
   (a) *Prioritize station* 1. *Assign as many servers as possible to station* 1 *with the caveat that unforced idling should be avoided* ($k = \max\{0, c_1 - i\}$).
   (b) *Prioritize station* 2. *Assign as many servers as possible to station* 2 *with the caveat that unforced idling should be avoided* ($k = \min\{c_1, j - c_2\}$).

PROOF: We show the results in the finite horizon case. The other cases follow similarly. The first result follows trivially from the fact that unforced idling is not optimal. This implies that for $j \leq c_2$ (there is no extra work for flexible servers to do), the optimal value for k is zero.

Assume now that $j > c_2$ and note that the no unforced idling assumption also rules out the possibility of an optimal k in (5.4) such that $k < c_1 - i$ and $k < j - c_2$ ($k < \min\{c_1 - i, j - c_2\}$); otherwise servers would idle at station 1 while there is work in the buffer at station 2. Similarly for $k > c_1 - i$ and $k > j - c_2$ ($k > \max\{c_1 - i, j - c_2\}$), we would idle servers at station 2 while there is work to do at station 1. This leaves two cases to consider. In the first case, $c_1 - i > j - c_2$; that is, there is more excess capacity at station 1 than there is work at station 2. Consider the following set:

$$B(i,j) := \{0 \leq k \leq c_1 \mid j - c_2 < k \leq c_1 - i\}.$$

For $i < c_1$ with $k \in B(i,j)$, the minimum in (5.4) is independent of k. Moreover, this is the same value we would obtain if $k = c_1 - i$. Thus, an optimal action is to set $k = c_1 - i$ and the second result is proven. Of course, when $i \geq c_1$, this case cannot occur (when $j > c_2$). Now consider the case where $c_1 - i \leq j - c_2$; that is, there are more customers at station 2 than can be served by the excess flexible servers at station 1. Define the following set:

$$A(i,j) := \{0 \leq k \leq c_1 \mid c_1 - i \leq k \leq j - c_2\}.$$

Thus,

$$\begin{aligned} v_{n+1}(i,j) = {} & ih_1 + jh_2 + \lambda_1(v_n(i+1,j) - v_n(i,j)) + \lambda_2(v_n(i,j+1) - v_n(i,j)) \\ & + g(i)[v_n(i-1,j+1) - v_n(i,j)] \\ & + \min_{k \in A(i,j)} \{-k\mu_1[v_n(i-1,j) - v_n(i,j)] + k\mu_3(v_n(i,j-1) - v_n(i,j))\} \\ & + c_1\mu_1(v_n(i-1,j) - v_n(i,j)) + c_2\mu_2(v_n(i,j-1) - v_n(i,j)) + v_n(i,j). \end{aligned}$$

Since the quantities in the minimum are linear in k, the optimal action is found at one of the extreme points. That is to say, the optimal action is at one of the following points: $k = \min\{c_1, j - c_2\}$ or $k = \max\{0, c_1 - i\}$. This proves the last case and the result is proven. ■

The previous theorem implies that when the queue lengths are large enough so that no servers are forced to idle, the prioritization of customer classes defines where all of

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

the servers should be placed; the servers are not split between stations. In this case, one might think of the servers as a single server with rates $\bar{\mu}_1 = c_1\mu_1$ or $\bar{\mu}_3 = c_1\mu_3$ and $\bar{\mu}_2 = c_2\mu_2$, respectively. With this observation it seems intuitive that the allocation decision can be translated to solving the analogous system with a single server at each station.

5.2. The Single-Server Case

Case 3 of Theorem 5.2 still begs the question "When should one prioritize station 1 or 2?" Suppose $c_1 = c_2 = 1$. To ease notation define $v_n(-1,j) = v_n(0,j)$ and $v_n(i,-1) = v_n(i,0)$. Note that the optimality equations in the finite horizon case reduce to

$$\begin{aligned}
&v_{n+1}(i,j)\\
&\quad = ih_1 + jh_2 + \theta\Big(\lambda_1 v_n(i+1,j) + \lambda_2 v_n(i,j+1) + g(i)v_n(i-1,j+1) + \mu_2 v_n(i,j-1)\\
&\qquad\qquad + \min_{k\in\{0,1\}}\Big\{(1-k)\mu_1 v_n(i-1,j) + k\mu_3 v_n(i,j-1)\\
&\qquad\qquad\qquad + [1 - (\lambda_1 + \lambda_2 + g(i) + (1-k)\mu_1\\
&\qquad\qquad\qquad\quad + k\mu_3 + \mu_2)]v_n(i,j)\Big\}\Big)\\
&\quad = ih_1 + jh_2 + \theta\Big(\lambda_1 v_n(i+1,j) + \lambda_2 v_n(i,j+1) + g(i)\beta v_n(i-1,j+1)\\
&\qquad\qquad + \mu_1 v_n(i-1,j) + \mu_2 v_n(i,j-1)\\
&\qquad\qquad + [1 - (\lambda_1 + \lambda_2 + g(i) + \mu_1 + \mu_2)]v_n(i,j)\\
&\qquad\qquad + \min_{k\in\{0,1\}}\Big\{k\mu_3[v_n(i,j-1) - v_n(i,j)]\\
&\qquad\qquad\qquad - k\mu_1[v_n(i-1,j) - v_n(i,j)]\Big\}\Big). \qquad (5.7)
\end{aligned}$$

From (5.7) note that it is optimal to serve at station 1 (2) when $\mu_3(v_n(i,j-1) - v_n(i,j)) \geq (\leq)\mu_1(v_n(i-1,j) - v_n(i,j))$. The next lemma follows in precisely the same manner as Proposition 3.1. The proof is omitted for brevity.

LEMMA 5.3: *Let $y = v_n, v_\theta$, or u depending on the optimality criterion. For all states (i,j) the following inequalities hold:*

1. $y(i,j+1) \geq y(i,j)$,
2. $y(i+1,j) \geq y(i,j)$,

for each θ, where in the finite horizon case the result holds for all n.

For any function f on the state space, let $\Delta_1 f(i,j) := f(i+1,j) - f(i,j)$ and $\Delta_2 f(i,j) := f(i,j+1) - f(i,j)$. Additionally, recall that $a_{k+1}b_{k+1} - a_k b_k = (a_{k+1} -$

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

$a_k)b_{k+1} + a_k(b_{k+1} - b_k) = a_{k+1}(b_{k+1} - b_k) + (a_{k+1} - a_k)b_k$. The next result states that when $\mu_1 = \mu_3$, there exists an optimal policy that is monotone in the number of customers at station 2. This implies that for each fixed i, there exists a (possibly infinite) control limit $L_2(i)$ such that it is optimal for the flexible server to work at station 1 for $j < L_2(i)$ and at station 2 for $j \geq L_2(i)$.

THEOREM 5.4: *Suppose $\mu_1 = \mu_3$ and let $y = v_n, v$, or u depending on the optimality criterion. The following hold:*

1. *Assume $\mu_2 \geq \mu_1$. For all $i,j \geq 1$, we have $y(i,j-1) - y(i-1,j) - [y(i,j) - y(i-1,j+1)] = \Delta_2 y(i-1,j) - \Delta_2 y(i,j-1) \geq 0$.*
2. *Suppose $\beta = 0$.*
   (a) *For all $i,j \geq 0$, we have $y(i+2,j) - y(i+1,j) - [y(i+1,j) - y(i,j)] = \Delta_1(\Delta_1)y(i,j) = \Delta_1^2 y(i,j) \geq 0$ (convexity in i).*
   (b) *For all $i \geq 0, j \geq 1$, we have $y(i+1,j) - y(i,j) - [y(i+1,j-1) - y(i,j-1)] = \Delta_1 y(i,j) - \Delta_1 y(i,j-1) \geq 0$.*
   (c) *For all $i,j \geq 1$, we have $y(i,j-1) - y(i-1,j) - [y(i+1,j-1) - y(i,j)] = \Delta_1 y(i-1,j) - \Delta_1 y(i,j-1) \leq 0$.*

   That is to say, from statements 1 and 2(c) we have the following under each of the optimality criteria:
3. *Assuming $\mu_2 \geq \mu_1$, if it is optimal to allocate the flexible server to station 2 in state (i,j), it is optimal to allocate it to station 2 for $(i,j+1)$ (monotone in j).*
4. *When $\beta = 0$, if it is optimal to allocate the flexible server to station 1 in state (i,j), it is optimal to allocate it to station 1 for $(i+1,j)$ (monotone in i).*

PROOF: We show the results for v_n. The others will follow by taking limits. Trivially, all of the results hold for $n = 0$. Suppose the results hold for n and consider $n+1$. A little algebra yields

$$\begin{aligned} v_{n+1}(i,j) = ih_1 + jh_2 + \theta\Big(&\lambda_1 v_n(i+1,j) + \lambda_2 v_n(i,j+1) + g(i)v_n(i-1,j+1) \\ &+ [1 - (\lambda_1 + \lambda_2 + g(i) + \mu_1 + \mu_2)]v_n(i,j) \\ &+ \mu_1 v_n(i-1,j) + \mu_2 v_n(i,j-1) \\ &+ \mu_1 \min_{k\in\{0,1\}} \{k[v_n(i,j-1) - v_n(i-1,j)]\}\Big). \end{aligned} \tag{5.8}$$

Let k be the optimal allocation (0 or 1) at time $n+1$ in state $(i-1,j+1)$ and let k' be the optimal allocation in state $(i,j-1)$. Using the same allocation in states $(i-1,j)$

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

and (i,j), respectively, we have for $i \geq 1, j \geq 2$ (suppressing the discount factor θ),

$$\begin{aligned}
&\Delta_2 v_{n+1}(i-1,j) - \Delta_2 v_{n+1}(i,j-1) \\
&\quad \geq \lambda_1[\Delta_2 v_n(i,j) - \Delta_2 v_n(i+1,j-1)] + \lambda_2[\Delta_2 v_n(i-1,j+1) - \Delta_2 v_n(i,j)] \\
&\qquad + \mu_1[\Delta_2 v_n(i-2,j) - \Delta_2 v_n(i-1,j-1)] \\
&\qquad + \mu_2[\Delta_2 v_n(i-1,j-1) - \Delta_2 v_n(i,j-2)] \\
&\qquad + k\mu_1[\Delta_2 v_n(i-1,j-1) - \Delta_2 v_n(i-2,j)] \\
&\qquad + k'\mu_1[\Delta_2 v_n(i-1,j-1) - \Delta_2 v_n(i,j-2)] \\
&\qquad + g(i-1)[\Delta_2 v_n(i-2,j+1) - \Delta_2 v_n(i-1,j)] \\
&\qquad + [1 - (\lambda_1 + \lambda_2 + \mu_1 + \mu_2 + g(i))][\Delta_2 v_n(i-1,j) - \Delta_2 v_n(i,j-1)],
\end{aligned}$$

where the inductive hypothesis applies in each case except for the term with coefficient k. In this case, if $k = 1$, the term cancels with the first term with coefficient μ_1. This is the case, for example, when $i = 1$. At $j = 1$, $k' = 0$, but $k = 0$ or 1 and we must allocate the server to station 1 in states $(i-1,1)$ and $(i,1)$. The changes in the terms with coefficient μ_1 and μ_2 (not including the uniformization terms which remain the same) are

$$\mu_1 \Delta_{2-k} v_n(i-2,1) + (\mu_2 - \mu_1)\Delta_2 v_n(i-1,0) \geq 0,$$

where the inequality follows by Proposition 3.1 and the assumption that $\mu_2 \geq \mu_1$. This completes the proof of statement 1.

Consider statement 2(a). Using (5.8) and ignoring the terms associated with $g(i)$ (since $\beta = 0$) yields

$$\begin{aligned}
\Delta_1^2 v_{n+1}(i,j) = \theta\Big(&\lambda_1 \Delta_1^2 v_n(i+1,j) + \lambda_2 \Delta_1^2 v_n(i,j+1) + [1 - (\lambda_1 + \lambda_2 + \mu_1 + \mu_2)] \\
&\times \Delta_1^2 v_n(i,j) + \mu_1 \Delta_1^2 v_n(i-1,j) + \mu_2 \Delta_1^2 v_n(i,j-1) \\
&+ \mu_1\Big(\min\{v_n(i+2,j-1) - v_n(i+1,j), 0\} \\
&\quad - \min\{v_n(i+1,j-1) - v_n(i,j), 0\} \\
&\quad - \min\{v_n(i+1,j-1) - v_n(i,j), 0\} \\
&\quad + \min\{v_n(i,j-1) - v_n(i-1,j), 0\}\Big)\Big).
\end{aligned}$$

It should be clear that the terms associated with λ_1, λ_2, μ_2, and uniformization all are nonnegative by the inductive hypothesis. Consider the terms with coefficient μ_1. Let k_1 (0 or 1) be the optimal allocation in state $(i+2,j)$ and k_1' be the optimal allocation in state (i,j). Note that the inductive hypothesis from statement 2(c) implies that the optimal policy is monotone in i, so $k_1' = 0$ implies $k_1 = 0$. Assuming that we use k_1

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

in the second minimum and k_1' in the third, we get a lower bound on the terms in question of

$$\mathcal{A}(i,j) := \mu_1 \Delta_1^2 v_n(i-1,j) + \mu_1(k_1[\Delta_1 v_n(i+1,j-1) - \Delta_1 v_n(i,j)] - k_1'[\Delta_1 v_n(i,j-1) - \Delta_1 v_n(i-1,j)]). \quad \textbf{(5.9)}$$

There are three cases to consider. If $k_1 = k_1' = 1$, then $\mathcal{A}(i,j) = \mu_1 \Delta_1^2 v_n(i,j-1)$. If $k_1 = k_1' = 0$, then $\mathcal{A}(i,j) = \mu_1 \Delta_1^2 v_n(i-1,j)$. In either case, the inductive hypothesis yields 2(a). If $k_1 = 0$ and $k_1' = 1$, then a little algebra yields

$$\mathcal{A}(i,j) = \mu[\Delta_1 v_n(i,j) - \Delta_1 v_n(i,j-1)] \geq 0,$$

where the inequality holds by the inductive assumption 2(b).

To show that statement 2(b) holds, consider only the terms of $\Delta_1 v_{n+1}(i,j) - \Delta_1 v_{n+1}(i,j-1)$ that have coefficient μ_1 (the other terms are analogous to the previous part):

$$\begin{aligned}
&\mu_1[\Delta_1 v_n(i-1,j) - \Delta_1 v_n(i-1,j-1)] \\
&\quad + \mu_1\Big(\min\{v_n(i+1,j-1) - v_n(i,j), 0\} - \min\{v_n(i,j-1) - v_n(i-1,j), 0\} \\
&\qquad - [\min\{v_n(i+1,j-2) - v_n(i,j-1), 0\} \\
&\qquad - \min\{v_n(i,j-2) - v_n(i-1,j-1), 0\}]\Big).
\end{aligned}$$

Suppose k_2 and k_2' represent the optimal allocations in states $(i+1,j)$ and $(i,j-1)$, respectively. If we mimic the allocations in the other two minimums, we have the following lower bound:

$$\begin{aligned}
\mathcal{B}(i,j) := {} & \mu_1[\Delta_1 v_n(i-1,j) - \Delta_1 v_n(i-1,j-1)] \\
& + k_2 \mu_1[\Delta_1 v_n(i,j-1) - \Delta_1 v_n(i-1,j)] \\
& - k_2' \mu_1[\Delta_1 v_n(i,j-2) - \Delta_1 v_n(i-1,j-1)].
\end{aligned}$$

If $k_2 = k_2' = 1$, then $\mathcal{B}(i,j) = \mu_1[\Delta_1 v_n(i,j-1) - \Delta_1 v_n(i,j-2)] \geq 0$, where the inequality follows from the inductive hypothesis of statement 2(b). If $k_2 = k_2' = 0$, then, $\mathcal{B}(i,j) = \mu_1[\Delta_1 v_n(i-1,j) - \Delta_1 v_n(i-1,j-1)] \geq 0$, again by the inductive hypothesis. Consider now the case when $k_2 = 1$ and $k_2' = 0$. In this case, $\mathcal{B}(i,j) = \mu_1[\Delta_1^2 v_n(i-1,j-1)] \geq 0$. To complete the proof of statement 2(b), suppose $k_2 = 0$ and $k_2' = 1$. In this case, we have $\mathcal{B}(i,j) = \mu_1[\Delta_1 v_n(i-1,j) - \Delta_1 v_n(i,j-1)] + \mu_1[\Delta_1 v_n(i,j-1) - \Delta_1 v_n(i,j-2)] \geq 0$, by the inductive hypothesis.

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

Consider now statement 2(c). Again we note that only the terms with coefficient μ_1 present a challenge. These terms are

$$\begin{aligned}
&\mu_1[\Delta_1 v_n(i-2,j) - \Delta_1 v_n(i-1,j-1)] + \mu[\Delta_1 v_n(i-1,j-1) - \Delta_1 v_n(i,j-2)] \\
&\quad + \mu_1\Big(\min\{v_n(i,j-1) - v_n(i-1,j), 0\} - \min\{v_n(i-1,j-1) - v_n(i-2,j), 0\} \\
&\qquad - \min\{v_n(i+1,j-2) - v_n(i,j-1), 0\} \\
&\qquad + \min\{v_n(i,j-2) - v_n(i-1,j-1), 0\}\Big).
\end{aligned}$$

Let k_3 and k_3' represent the optimal allocation in states $(i-1,j)$ and $(i+1,j-1)$, respectively. An upper bound on the terms with coefficient μ_1 is

$$\begin{aligned}
\mathcal{C}(i,j) := \mu_1[\Delta_1 v_n(i-2,j) - \Delta_1 v_n(i-1,j-1)] \\
+ k_3\mu_1[\Delta_1 v_n(i-1,j-1) - \Delta_1 v_n(i-2,j)] \\
- k_3'\mu_1[\Delta_1 v_n(i,j-2) - \Delta_1 v_n(i-1,j-1)].
\end{aligned}$$

If $k_3 = k_3' = 1$, then $\mathcal{C}(i,j) = \mu_1[\Delta_1 v_n(i-1,j-1) - \Delta_1 v_n(i,j-2)] \le 0$, where the inequality holds from the inductive hypothesis of statement 2(c) when $j \ge 2$ and statement 2(a) when $j = 1$. Similarly, if $k_3 = k_3' = 0$, then $\mathcal{C}(i,j) = \mu_1[\Delta_1 v_n(i-2,j) - \Delta_1 v_n(i-1,j-1)] \le 0$, where again the inductive hypothesis yields the inequality when $i \ge 2$ and Lemma 5.3 admits that it holds when $i = 1$. If $k_3 = 0$ and $k_3' = 1$, then

$$\begin{aligned}
\mathcal{C}(i,j) = \mu_1[\Delta_1 v_n(i-2,j) - \Delta_1 v_n(i-1,j-1)] \\
+ \mu_1[\Delta_1 v_n(i-1,j-1) - \Delta_1 v_n(i,j-2)] \le 0,
\end{aligned}$$

where, again, the inductive hypothesis yields the inequality. If $k_3 = 1$ and $k_3' = 0$, then $\mathcal{C}(i,j) = 0$ (this case does not occur when $j = 1$). ■

THEOREM 5.5: *Suppose* $\mu_1 h_1 \ge \mu_3 h_2$ *and let* $y = v_n, v,$ *or* u *depending on the optimality criterion. If*

1. $\beta = 0$ *(no upgrades)*

or

2. $\mu_1 \ge \mu_3$,

then for all $i,j \ge 0$, *we have*

$$\begin{aligned}
-\mu_3\Delta_2 y(i,j-1) &= \mu_3[y(i,j-1) - y(i,j)] \\
&\ge \mu_1[y(i-1,j) - y(i,j)] \\
&= -\mu_1\Delta_1 y(i-1,j),
\end{aligned}$$

that is, $\mu_1\Delta_1 y(i-1,j) \ge \mu_3\Delta_2 y(i,j-1)$. *Thus, under each of the optimality criteria, there exists an optimal policy that always serves at station 1, except to avoid unforced idling.*

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

PROOF: Consider the finite horizon case. The result holds trivially for $n = 0$. Suppose it holds for n. This implies that it is optimal to place the flexible server at station 1 except to avoid unforced idling. Consider $n + 1$ and suppose $i \geq 1, j \geq 1$:

$$\begin{aligned}
&\Delta_1 v_{n+1}(i-1,j) \\
&\quad = h_1 + \theta\Big[\lambda_1 \Delta_1 v_n(i,j) + \lambda_2 \Delta_1 v_n(i-1,j+1) \\
&\qquad\qquad + \mu_1 \Delta_1 v_n(i-2,j) + \mu_2 \Delta_1 v_n(i-1,j-1) \\
&\qquad\qquad + g(i) v_n(i-1,j+1) - g(i-1) v_n(i-2,j+1) \\
&\qquad\qquad + [1 - (\lambda_1 + \lambda_2 + \mu_1 + \mu_2)] \Delta_1 v_{n+1}(i-1,j) \\
&\qquad\qquad - g(i) v_n(i,j) + g(i-1) v_n(i-1,j)\Big].
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\Delta_2 v_{n+1}(i,j-1) \\
&\quad = h_2 + \theta\Big[\lambda_1 \Delta_2 v_n(i+1,j-1) + \lambda_2 \Delta_2 v_n(i,j) \\
&\qquad\qquad + \mu_1 \Delta_2 v_n(i-1,j-1) + \mu_2 \Delta_2 v_n(i,j-2) + g(i) \Delta_2 v_n(i-1,j) \\
&\qquad\qquad + [1 - (\lambda_1 + \lambda_2 + \mu_1 + \mu_2 + g(i))] \Delta_2 v_n(i,j-1)\Big].
\end{aligned}$$

Consider now $\mu_1 \Delta_1 v_{n+1}(i-1,j) - \mu_3 \Delta_2 v_{n+1}(i,j-1)$. Suppose $g(i) = 0$ for all i. Note that the differences in the terms involving λ_1, λ_2, μ_1, and μ_2 and that due to uniformization are nonnegative by the inductive hypothesis in each case. That is to say, the result holds for $\beta = 0$ as desired. Suppose now that $g(i) > 0$ for some i. Combining the remaining terms yields

$$\begin{aligned}
&(1 - (\lambda_1 + \lambda_2 + \mu_1 + \mu_2 + g(i)))\Big[\mu_1 \Delta_1 v_n(i-1,j) - \mu_3 \Delta_2 v_n(i,j-1)\Big] \\
&\quad + g(i)[\mu_1 \Delta_1 v_n(i-2,j+1) - \mu_3 \Delta_2 v_n(i-1,j)] \\
&\quad + \Delta_1 g(i-1) \mu_1 (v_n(i-2,j+1) - v_n(i-1,j)) \\
&\qquad = (1 - (\lambda_1 + \lambda_2 + \mu_1 + \mu_2 + g(i)))\Big[\mu_1 \Delta_1 v_n(i-1,j) - \mu_3 \Delta_2 v_n(i,j-1)\Big] \\
&\qquad\quad + g(i-1)[\mu_1 \Delta_1 v_n(i-2,j+1) - \mu_3 \Delta_2 v_n(i-1,j)] \\
&\qquad\quad + \Delta_1 g(i-1)(\mu_1 - \mu_3)(\Delta_2 v_n(i-1,j)).
\end{aligned}$$

The second result now holds for $i \geq 2$ by the inductive hypothesis, the assumption that $\mu_1 \geq \mu_3$, and Lemma 5.3. It remains to consider the case with $i = 1$. In this case, the optimal policy in $(i-1,j)$ is to assign both servers to station 2 (or not at all when

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

$j = 1$). Thus,

$$\begin{aligned}
&\Delta_1 v_{n+1}(0,j) \\
&\quad = h_1 + \theta\Big[\lambda_1 \Delta_1 v_n(1,j) + \lambda_2 \Delta_1 v_n(0,j+1) \\
&\qquad\qquad + \mu_2 \Delta_1 v_n(0,j-1) + [1 - (\lambda_1 + \lambda_2 + \mu_1 + \mu_2)]\Delta_1 v_n(0,j) \\
&\qquad\qquad + \mu_3 \Delta_2 v_n(0,j-1)\Big].
\end{aligned}$$

Again, the terms of $\mu_1 \Delta_1 v_{n+1}(0,j) - \mu_3 \Delta_2 v_{n+1}(1,j-1)$ associated with λ_1, λ_2, and uniformization (recall $g(1) = 0$) are nonnegative via the inductive hypothesis; similarly for $\mu_2[\mu_1 \Delta_1 v_n(0,j-1) - \mu_3 \Delta_2 v_n(1,j-2)] \geq 0$. The last term $\Delta_2 v_n(0,j-1)$ cancels and the proof is complete for all n. Taking limits as $n \to \infty$ yields the analogous result in the infinite horizon case. To get the average case, define a subsequence $\{\theta_n, n \geq 0\}$ as in Proposition 5.1, part 2. Taking limits and applying the ACOEs yields the results in the average case. ■

6. NUMERICAL STUDY AND EXAMPLES

In this section we provide several examples and a detailed numerical study. Unless otherwise stated, in each of the examples the state space was truncated at 50 for each station to make the Markov decision process tractable. Given the results of Theorem 5.5, one might wonder if the condition that $\mu_1 \geq \mu_3$ plays a significant role in determining the optimal policy. The next example shows that indeed it does.

Example 6.1: Assume $c_1 = c_2 = 1$ and suppose $\lambda_1 = 2, \lambda_2 = 3, \mu_1 = 2, \mu_2 = \mu_3 = 3$, $\beta = 1, L = 7, h_1 = 1.5$, and $h_2 = 1$.

Since $\mu_1 h_1 = \mu_3 h_2 = 3$, one might conjecture that the c-μ rule implies that the optimal policy is to leave the flexible server at station 1 except to avoid unforced idling. However, it turns out that the optimal policy is to have the flexible server serve work at station 2 except to avoid unforced idling. Upon further inspection, this is intuitive since $\lambda_1 < L\beta$, so that on the fluid scale, all of the work is pushed to station 2. On the other hand, if we replace μ_1 with 3, then, as expected, the optimal policy is to leave the flexible server at station 1 except to avoid unforced idling (as expected from the results of Theorem 5.5). In the prior case, the average cost is 7.8077, whereas in the latter case, the average cost is 6.0735 (a 22% difference). Moreover, if we use the policy that places the flexible server at station 1, in the system with $\mu_1 = 2$ the average cost is 16.4862 (a 111% increase!).

The next example shows that there are cases that warrant a search for an *optimal* upgrade rate.

Example 6.2: Suppose we have the following parameters: $\lambda_1 = 3$, $\lambda_2 = 1$, $\mu_1 = 3$, $\mu_2 = \mu_3 = 1.5, h_1 = 1, h_2 = 7$, and $L = 50$.

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

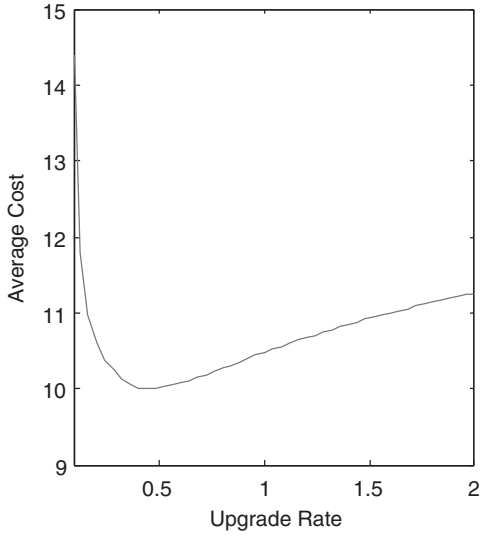

FIGURE 2. Average cost versus upgrade rate.

Allowing β to range from zero to 2 yields the results shown in Figure 2. It should be clear that when $\beta = 0$, the system is unstable ($Q_1(t) \to \infty$). Moreover, one should note that the curve is unimodal but not convex in β. Thus, given a fixed set of parameters, one might search to find the best upgrade rate.

An experiment was designed to see how changing the system utilization and the holding costs affect the usefulness of the optimal allocation of the flexible server in a system with upgrades. In each run, we let the maximum buffer size of each queue be 35. We also fix the following parameters: $\lambda_1 = \lambda_2 = 1$, $L = 10$, and $\beta = 0.05$ and let $\mu_2 = \mu_3$. Define

$$\rho_1 = \frac{\lambda_1}{\mu_1 + L\beta},$$
$$\rho_2 = \frac{\lambda_2 + L\beta}{\mu_2(2 - (\lambda_1 - L\beta)/\mu_1)}.$$

We note that in neither case does $\rho_i < 1$ for $i = 1, 2$ guarantee stability of the system. However, each in some sense gives a measure of the congestion at each station. We allow $\rho_1, \rho_2 \in \{.6, .8, .95\}$ and $h_1, h_2 \in \{0.5, 1, 3, 7\}$. The service rates implied by each combination of ρ_1 and ρ_2 are displayed in Table 1.

The average costs of the optimal policy and the base policy that assigns the flexible server to station 1 except to avoid unforced idling were computed. For each run, the percent difference (divided by 100) is obtained. The average cost percent difference between the optimal and base policies across utilization and holding costs combinations are displayed in Table 2. One immediately notices that as ρ_2 increases, the

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

TABLE 1. Service Rate Inputs

μ_1		ρ_1		
μ_2		.6	.8	.95
ρ_2	.6	1.1667	0.7500	0.5526
		1.5909	1.8750	2.2826
	.8	1.1667	0.7500	0.5526
		1.1932	1.4063	1.7120
	.95	1.1667	0.7500	0.5526
		1.0048	1.1842	1.4416

TABLE 2. (Average Percent Difference)/100 Across Utilization and Holding Costs

		ρ_1		
		.6	.8	.95
ρ_2	.6	0.0659	0.0579	0.0205
	.8	0.0995	0.2242	0.1125
	.95	0.1019	0.3270	0.3254

		h_1			
		.5	1	3	7
h_2	0.5	0.0845	0.0310	0.0053	0.0029
	1	0.1788	0.0845	0.0147	0.0046
	3	0.3833	0.2548	0.0845	0.0237
	7	0.5186	0.4102	0.2073	0.0845

average cost difference increases. This stands to reason since the higher the congestion in station 2, the more important it becomes to spend time away from station 1 (unlike the base policy). By the same token, the optimal policy is not often different than the base policy when ρ_2 is small and ρ_1 is high (it is optimal to serve at station 1). The point should also be made that the difference is significant when $\rho_2 = .95$; it is above 10% in each case and above 32% when $\rho_1 \geq .8$.

We see similar results when we consider the average difference with regard to holding costs. Of course, when the holding costs are the same and $\mu_1 \geq \mu_2$, Theorem 5.5 implies that the base case is optimal. This is only violated in one instance (see Table 1), so that the terms down the diagonal are the same. As for the other terms, one should note that as h_2 increases relative to h_1 the difference in costs becomes more pronounced. In the worst case, it is over 17%.

We also examined (via simulation) the differences in average waiting times between the base case and the optimal policy. To do so, we calculated the waiting times of customers of each class after 30,000 time units of burn-in and 30,000 actual time units. This was done for 30 replications over each of the combinations of the previous study. The results are contained in Table 3, where W_i represents the average

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

Table 3. (Average Percent Decrease in Average Waiting Times)/100 Across Utilization and Holding Costs

W_1 / W_2		ρ_1 = .6	ρ_1 = .8	ρ_1 = .95
ρ_2	.6	−0.1707	−0.0931	−0.0478
		0.1750	0.2623	0.2432
	.8	−0.3884	−0.1083	−0.0317
		0.2347	0.5901	0.5735
	.95	−0.2293	0.7544	0.7907
		0.0502	0.8163	0.9800

W_1 / W_2		h_1 = .5	h_1 = 1	h_1 = 3	h_1 = 7
h_2	0.5	0.1690	0.0643	−0.2044	−0.3264
	0.5	0.4476	0.5606	0.6652	0.6764
	1	0.1992	0.1645	−0.0132	−0.2486
	1	0.3628	0.4459	0.6097	0.6711
	3	0.0879	0.1980	0.1661	0.0248
	3	0.1326	0.3278	0.4460	0.5762
	7	0.1057	0.0945	0.2001	0.1642
	7	0.1279	0.1323	0.3514	0.4446

waiting time of type *i* customers and type 1 customers are called type 1 regardless of which server actually serves them. The upper table, which for a fixed load averages over the 16 holding cost cases, shows that for moderate loads, we get the intuitive result that the optimal policy has relatively shorter waiting times for type 2 customers and longer waiting times for type 1 customers. However, at high loads, there is a dramatic improvement for *both* types. This appears to be due to the fact that more customers are upgraded than at lower loads, and the base policy does poorly with respect to these. In bottom table, we see that for almost all holding cost pairs, there is improvement for both types.

7. CONCLUSIONS

In this article we have considered a new method of mitigating the concerns of waiting times for low-priority customers while maintaining the prioritization in a system of two parallel stations and flexible servers. This is obtained by *upgrading* customers to high priority if they have been waiting in queue for some time. This is done using a simple mechanism that is analytically tractable. From a managerial standpoint, the effectiveness of such a simple mechanism suggests that allowing limited upgrades can effectively take advantage of unused capacity.

https://doi.org/10.1017/S0269964809990222 Published online by Cambridge University Press

We also showed several results that at first glance might seem counterintuitive. In particular, we showed that the allocation of flexible servers in this system needs to consider the relationship between μ_1 and μ_3 directly (not just $\mu_1 h_1$ and $\mu_3 h_2$). This occurs in both the stability calculations and the calculation of optimal policies. Moreover, we provide conditions under which the optimal policy mirrors closely those that can be found in the literature.

We feel that this article opens the door for several areas of research. Since upgrades might be thought of as reneging without the customer actually leaving the system, it would be interesting to know how control would work in larger multiclass networks with reneging and upgrades. As we have pointed out, the questions of both stability and control are worthwhile. There is also the question of direct control of upgrades and flexible servers simultaneously. This is particularly interesting since one might consider both upgrades and downgrades; it is unclear how much is gained by providing immediate service in lieu of delayed service of lower-class customers. Since upgrades in effect control the arrival process to the high-priority queue, one can also ask the question if it would be more prudent to affect the arrival process by alternative means for example, by pricing or even admission control.

Acknowledgments

The work of the first author is supported by the Natural Sciences and Engineering Research Council of Canada. The work of the second author is partially supported by the National Science Foundation under Grant Nos. CMMI-0540808 and CMMI-0826255.

References

1. Ahn, H., Duenyas, I., & Zhang, R. (2004). Optimal control of a flexible server. *Advances In Applied Probability* 36(1): 139–170.
2. Aksin, Z., Armony, M., & Mehrotra, V. (2007). The modern call-center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* 16: 665–668.
3. Armony, M. & Maglaras, C. (2004). Contact centers with a call-back option and real-time delay information. *Operations Research* 52(4): 527–545.
4. Armony, M. & Maglaras, C. (2004). On customer contact centers with a call-back option: Customer decisions, routing rules and system design. *Operations Research* 52(2): 271–292.
5. Armony, M., Shimkin, N., & Whitt, W. (2007). The impact of delay announcements in many-server queues with abandonment. Preprint.
6. Bell, S. & Williams, R. (2001). Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Annals of Applied Probability* 11(3): 608–649.
7. Brémaud, P. (1981). *Point processes and queues. Martingale dynamics*. New York: Springer-Verlag.
8. Buyukkoc, C., Varaiya, P., & Walrand, J. (1985). The $c\mu$-rule revisited. *Advances in Applied Probability* 17(1): 237–238.
9. Dai, J. (1999). *Stability of fluid and stochastic processing networks*. Princeton, NJ: Centre for Mathematical Physics and Stochastics.
10. Dai, J. & Meyn, S. (1995). Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Transactions on Automatic Control* 40: 1889–1904.
11. Dai, J. & Tezcan, T. (2008). Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems* 59: 95–134.

12. Dai, J. & Tezcan, T. (in press). Dynamic control of n-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Operations Research*.
13. Gans, N., Koole, G. & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review and research prospects. *Manufacturing and Service Operations Management* 5(2): 79–141.
14. Gurvich, I. & Whitt, W. (2009). Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing and Service Operations Management* 11(2): 237–253.
15. Harrison, J. (1998). Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete-review policies. *Annals of Applied Probability* 8(3): 822–848.
16. Lippman, S. (1975). Applying a new device in the optimization of exponential queueing system. *Operations Research* 23(4): 687–710.
17. Mandelbaum, A. & Stolyar, A. (2004). Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research* 52(6): 836–855.
18. Sennott, L.I. (1999). *Stochastic dynamic programming and the control of queueing systems*. Wiley Series in Probability and Statistics. New York: Wiley.
19. Stanford, D.A. & Grassman, W.K. (1993). The bilingual server system: A queuing model featuring fully and partially qualified servers. *INFOR* 31(4): 261–278.
20. Ward, A.R. & Glynn, P.W. (2003). A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* 43(1): 103–128.
21. Ward, A.R. & Glynn, P.W. (2005). A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queueing Systems* 50(4): 371–400.