Probability in the Engineering and Informational Sciences (2023), **37**:3 833–851
doi:10.1017/S0269964822000213

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Scheduling servers in a two-stage queue with abandonments and costs

Gabriel Zayas-Cabán[1] and Amy L. Cochran[2]

[1] Industrial and Systems Engineering, BerbeeWalsh Department of Emergency Medicine, University of Wisconsin, Madison, WI, USA. E-mail: zayascaban@wisc.edu
[2] Department of Mathematics, Population Health Sciences, University of Wisconsin, Madison, WI, USA.

Keywords: Operations research, Queueing theory, Stochastic dynamic programming, Stochastic modeling

Abstract
We consider the assignment of servers to two phases of service in a two-stage tandem queueing system when customers can abandon from each stage of service. New jobs arrive at both stations. Jobs arriving at station 1 may go through both phases of service and jobs arriving at station 2 may go through only one phase of service. Stage-dependent holding and lump-sum abandonment costs are incurred. Continuous-time Markov decision process formulations are developed that minimize discounted expected and long-run average costs. Because uniformization is not possible, we use the continuous-time framework and sample path arguments to analyze control policies. Our main results are conditions under which priority rules are optimal for the single-server model. We then propose and evaluate threshold policies for allocating one or more servers between the two stages in a numerical study. These policies prioritize a phase of service before "switching" to the other phase when total congestion exceeds a certain number. Results provide insight into how to adjust the switching rule to significantly reduce costs for specific input parameters as well as more general multi-server situations when neither preemption or abandonments are allowed during service and service and abandonment times are not exponential.

1. Introduction

Many hospital systems (e.g., Lutheran Medical Center and UW Health) have implemented interventions known as "split flow" models to improve patient flow in the Emergency Departments (EDs) (cf., [21,29,49,50]). Unlike a typical ED, a split flow model stations an advanced practice provider (APP) rather than a nurse at patient intake (i.e., before or during triage). This provider briefly sees all walk-in patients and may initiate the care of all patients by placing lab, imaging, and medication orders (i.e., phase one service). The provider then stratifies patients, keeping those who do not require a traditional ED bed in a fast track or similar area (i.e., phase two service) and moving the rest to a queue for a traditional bed. Importantly, the same care provider is continuously switching between both phases of service, and while patients rarely leave before being seen while awaiting phase one service (e.g., triage), some may abandon the system before receiving final treatment.

Motivated by the ED split flow model, we consider the assignment of servers between two phases of service in a two-stage tandem queue. This and related models have been studied in the case when one or more job classes require service at multiple service stations in tandem by one or more servers (cf., [1–8,15,17,18,20,25,27,35,37,41,42,46,49]). Servers decide where to allocate efforts to optimize performance criteria on quality of service and/or congestion. Although widely studied, many of these models assume that jobs have unlimited patience and thus willing to wait indefinitely. However, there are practical and important situations, such as the ED, where jobs abandon the system before service.

Allowing for abandonments, we analyze this scheduling problem to find new control strategies that generalize to broader situations.

In the present study, we assume Poisson arrival streams to each phase of service with infinite waiting capacity, and one or more flexible servers that can serve arriving jobs having service times that are identically distributed and are independent, and are independent of inter-arrival times. We assume that no setup is required for the server to switch from processing jobs at one station to another. In addition, we assume that jobs can abandon from each phase of service. The goal is to provide server assignment policies to minimized expected discounted or average costs over an infinite horizon where each job class has linear holding costs and lump-sum abandonment costs.

To analyze this decision-making scenario, we first assume exponentially distributed service and abandonment times and use Markov decision processes (MDPs). For the multi-server model, we show that costs are nondecreasing in the number of customers in each phase of service, and that there exists a non-idling policy that is optimal so long as service preemption is allowed. We then show that if the number of customers at each queue exceeds the number of servers, there exists an optimal control policy that does not split the servers so long as service preemption is allowed. We then consider the single-server model and identify a set of conditions under which static priority rules are optimal. Outside these conditions, the optimal policy may be a complicated state-dependent policy. We thus consider a class of threshold policies which we evaluate in a two-part simulation study.

In the first part, we consider the single-server Markovian model and compare the performance of these threshold policies to the priority rules in the cases where we have proved that the priority rules are optimal. This is done to benchmark the proposed threshold policies. We then compare their relative performance in a discrete-event simulation for a wider range of parameter values. In the second part, we evaluate these threshold policies when abandonments during service and preemption are not allowed and service and abandonment times are no longer exponential.

The rest of the paper is organized as follows: Section 2 contains a summary of the literature related to our work. Section 3 describes in detail the model we consider, the MDP formulation of the server allocation problem, and some preliminary results that will be used throughout the paper. Section 4 contains our main analysis of static priority rules. In Section 5, we propose several allocation policies based on our analytical results and compare them numerically. We conclude with a discussion in Section 6.

2. Brief summary of the related literature

This study lies at the intersection of three areas: scheduling in tandem queueing systems, performance analysis of service policies in tandem queues, and MDPs with unbounded transition rates.

To our knowledge, the only other studies combining these three areas are [49,50]. In Zayas-Cabán *et al.* [49], policies for a two-stage tandem queueing model with abandonments and rewards accrued after service completion were analyzed using a continuous-time Markov decision process (CTMDP), motivated by the Triage and Treat Release Program at the Lutheran Medical Center in New York. In a follow-up study, Zayas-Cabán *et al.* [50] consider the same two-phase stochastic service system but where customers may only abandon the second phase of service. They introduce a class of policies they term K-level threshold policies, which prioritize phase 2 service unless there are K or more jobs in phase one service. Sufficient conditions are provided to ensure these policies yield a stable system. A heuristic is presented for choosing K in systems with abandonments. They analyze the performance of these heuristics in a simulation study. The present study differs from [49] and [50] by allowing arrivals to the second phase of service and by considering holding costs and abandonment costs (and not rewards). Holding and abandonment costs imply that cost rates are unbounded, which require stronger conditions to hold in order for the optimality equations to have a solution (cf., [13]). Furthermore, holding cost rates imply that stronger conditions are needed for the optimality of prioritizing station 2 to hold. The bounds for the value function that were obtained in [49] to imply that the optimality of static priority rules cannot be similarly obtained in the present study.

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

There is extensive literature on dynamic assignment of servers to different phases of service without abandonments (cf., [1–8,15,17,18,20,25,27,35–37,41,42,46,49]). Nelson [35], for instance, presents the optimality of priority rules akin to the classic c-μ rule but for a tandem system without abandonments in the context of a labor assignment problem over a finite horizon using the Pontryagin maximum principle. There is also extensive literature on performance analysis of single-server tandem queues without abandonments, which provide analysis techniques for determining stability conditions of different allocation policies such as priority rules (cf., [22–24,26,34,43]). We refer the reader to [49,50] for a recent review of this literature. More recently, Wang *et al.* [47] and Rastpour *et al.* [39] analyze multi-server Markovian queues with abandonments. Viewed as a level-dependent quasi-birth and death process, Wang *et al.* [47] analyze a multi-server tandem queue with abandonments where the second phase of service has a finite number of servers.

Lastly, we contend with a CTMDP with unbounded rates as a consequence of abandonments (cf., [10–14,16,28,32,33,38,40,44,45,48,49]). In particular, it extends the cost model considered by Down *et al.* [14] by allowing phase one service completions to join phase two service. Throughout the paper, and whenever possible, we compare our results to those obtained in Down *et al.* [14]. Although the remaining above-cited papers consider different models than the one considered here, they also provide approaches for how to analyze optimal controls for problems with unbounded rates. We remark that alternative approaches to CTMDPs have been used to contend with scheduling with abandonments (cf., [9,30,31]). For example, Atar *et al.* [9] consider the K-competing queues problem with many servers and introduce the $h\mu/\beta$ rule, which prioritizes the queue with the highest index $h_c\mu_c/\beta_c$, showing that it is asymptotically optimal in the so-called overloaded regime, as the number of servers tends to infinity.

3. Dynamics and control formulation

Suppose customers, or jobs, arrive to station 1 (2), or phase 1 (2) service, or queue 1 (2), of a tandem service system according to a Poisson process of rate λ_1 (λ_2) and immediately join the first (second) queue. Once the customer joins station 1 (2), their station 1 (2) patience time and service requirements are generated; the distribution of the former and latter will be first assumed to be exponential with rate β_1 (β_2) and μ_1 (μ_2), respectively. If the customer does not complete station 1 (2) service before the abandonment time ends, a lump-sum cost K_1 (K_2) is charged and the customer leaves the system without receiving service at station 1 or 2 (2). If the customer receives service at station 1 then, independently of the service time and arrival process, with probability $p \in [0, 1]$, the customer joins the queue at station 2. With probability $q := 1 - p$, the customer leaves the system forever.

There is a nonnegative holding cost h_c per job per unit time incurred for holding a customer in station $c \in \{1, 2\}$. There are $N \geq 1$ servers, each of which can be assigned to either station. We seek a non-anticipating policy that describes where to place the server based on the current state and potentially the history of states and actions taken. Within each station, the service discipline is first come first served (FCFS), and we consider both when service can and cannot be preempted and when abandonments of jobs in process are and are not allowed. Once a customer completes service at station 2, they leaves the system forever.

Fix a non-anticipating policy π, let $\{\sigma_n^\pi, n \geq 1\}$ denote the sequence of event times that includes arrivals, abandonments, and potential service completions, and let $\mathbb{Z}^+$ denote the set of nonnegative integers. For this analysis, we first assume that service can be preempted and abandonments during service are allowed. Throughout the paper, and whenever possible, we highlight when these restrictions can be relaxed. When service can be preempted and abandonments during service are allowed, the state space is $\mathbb{X} := \{(x_1, x_2) \,|\, x_2, x_2 \in \mathbb{Z}^+\}$, where x_1 (resp. x_2) denotes the number of customers at station 1 (resp. 2) and the set of available actions at state $x \in \mathbb{X}$ is $A(x) = \mathbb{A} := \{(y_1, y_2) \,|\, y_1, y_2 \in \mathbb{Z}^+, y_1 + y_2 \leq N\}$, where y_1 (y_2) represents the number of servers assigned to station 1 (2).[1] For $\alpha > 0$,

[1]When abandonments during service and preemption are not allowed, we also need to keep track of each server status so that the state space becomes $\mathbb{X} := \{(x_1, x_2, y_1, y_2) \,|\, x_1, x_2, y_1, y_2 \in \mathbb{Z}^+, 0 \leq y_1 + y_2 \leq N\}$ where x_1 (resp. x_2) denotes the number of customers at station 1 (resp. 2), and y_1 and y_2, respectively, denote the number of servers serving at station 1 and serving at station 2. Furthermore, decisions are made

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

the finite horizon (of length t), α-discounted expected cost for a non-anticipating policy π is given by $v^{\pi}_{t,\alpha}(x) \equiv \mathbb{E}^{\pi}_{x}[\sum_{n=0}^{N(t,\pi)} e^{-\alpha\sigma_n} k(X_n, a_n)] + \int_0^t e^{-\alpha s}\mathbb{E}^{\pi}_{x}[\sum_{c=1}^{2} h_c Q^{\pi}_{c}(s))]\, ds$, where $Q^{\pi}_{c}(s)$ denotes the customer class $c \in \{1,2\}$ queue length process at time $s \geq 0$, and X_n and a_n represent, respectively, the state of the system and the type of event seen at the time of the nth decision. The function $k(\cdot,\cdot)$ denotes the fixed cost; that is, if σ_n denotes a station $c \in \{1,2\}$ customer abandonment, then $k(X_n, a_n) = K_c$ and it is zero otherwise. For fixed $x \in \mathbb{X}$, the infinite horizon discounted expected cost under policy π is $v^{\pi}_{\alpha}(x) \equiv \lim_{t\to\infty} v^{\pi}_{t,\alpha}(x)$. The long-run average cost rate is $\rho^{\pi}(x) \equiv \limsup_{t\to\infty} v^{\pi}_{t,0}(x)/t$.

4. Dynamic control

We present two results that are used throughout when abandonments during service and preemption are allowed. The first is the monotonicity of the value functions. The second says that there is an optimal policy that does not idle the servers whenever there are customers waiting so long as service can be preempted. The latter is used to simplify the optimality equations. In the interest of brevity, we omit the proofs.

Proposition 4.1. *The following hold:*

1. *For all* $x = (x_1, x_2) \in \mathbb{X}$
$$v_{\alpha}(x + e_c) \geq v_{\alpha}(x), \quad c = 1, 2$$
where e_c *denotes the* c*th standard basis vector in* $\mathbb{R}^2$ $(c = 1, 2)$*. Similarly, if* (g, w) *is a solution to the average cost optimality equations (defined below), the above statements hold with* v_{α} *replaced with* w*.*
2. *Under both the* α*-discounted cost and the average cost criterion, there exists a (Markovian) non-idling policy that is optimal.*

We remark that a similar result to the first part of Proposition 4.1 holds when abandonments during service and service preemption are *not* allowed, but a similar result to the second part of Proposition 4.1 only holds when abandonments during service are not allowed. When service preemption is not allowed, it may be optimal for the server to stay idle.

4.1. Optimality equations

For functions $f : \mathbb{X} \to \mathbb{R}$, let $f(x_1, x_2) := f((x_1, x_2))$ for $(x_1, x_2) \in \mathbb{X}$. The operator T acting on functions $f : \mathbb{X} \to \mathbb{R}$ is defined as follows.

$$\begin{aligned} Tf(x_1,x_2) := {} & x_1(h_1 + \beta_1 K_1) + x_2(h_2 + \beta_2 K_2) + \lambda_1 f(x_1+1, x_2) + \lambda_2 f(x_1, x_2+1) \\ & + x_1\beta_1 f(x_1-1, x_2) + x_2\beta_2 f(x_1, x_2-1) \\ & - (\lambda_1 + \lambda_2 + x_1\beta_1 + x_2\beta_2) f(x_1, x_2) \\ & + \min_{a_1 \in \{0,1,\ldots,N\}} \{\min\{x_1, a_1\}\mu_1[p f(x_1-1, x_2+1) + q f(x_1-1, x_2) - f(x_1, x_2)] \\ & + \min\{x_2, N - a_1\}\mu_2[f(x_1, x_2-1) - f(x_1, x_2)]\}. \end{aligned}$$

Theorem 4.2. *For any* $\alpha > 0$*, the following statements hold.*

1. *The value function* v_{α} *satisfies the discounted cost optimality equations (DCOE), that is,*
$$\alpha v_{\alpha}(x_1, x_2) = T v_{\alpha}(x_1, x_2), \quad (x_1, x_2) \in \mathbb{X}.$$

after service completions and the set of available actions at state $x \in \mathbb{X}$ becomes $A(x) := \{(a_1, a_2) \mid a_1, a_2 \in Z^+, a_1 + a_2 \leq N - y_1 - y_2\}$, where a_1 (a_2) represents the number of idling servers assigned to station 1 (2).

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

2. *There exists a deterministic stationary α-optimal policy.*
3. *Any policy satisfying the maximum in the DCOE is α-optimal.*

Proof. See Online Appendix in Supplementary material. □

Theorem 4.3. *If any of the following mutually exclusive conditions holds*

- $\min\{\beta_1, \beta_2\} > 0$*; or*
- $\beta_1 > 0$, $\beta_2 = 0$, *and if* $\lambda_2/\mu_2 < 1$ *for the multi-server model, or, if* $\lambda_1 \cdot (1/\pi_0(\mu_1 + \beta_1) + p(1 - P(Ab)/\mu_2) + \lambda_2/\mu_2 < 1$ *for the single-server model, where* π_0 *is long-run fraction of time that station 2 is empty under the non-idling policy that prioritizes station 2 when* $\beta_2 = 0$ *and* $P(Ab) = \beta_1/(\mu_1 + \beta_1)$ *; or*
- $\beta_1 = 0$, $\beta_2 > 0$ *and* $\lambda_1/\mu_1 < 1$,

then the following hold:

1. *There is a constant* $g \in \mathbb{R}$ *and a function* $w : \mathbb{X} \to \mathbb{R}$ *that satisfy the average cost optimality equations (ACOE), that is,*
$$g = Tw(x_1, x_2), \quad (x_1, x_2) \in \mathbb{X}.$$
2. *There exists a deterministic stationary average-optimal policy.*
3. *Any policy satisfying the maximum in the AROE is average-optimal, and* $g = \rho(x_1, x_2)$ *for all* $(x_1, x_2) \in \mathbb{X}$.

Proof. See Online Appendix in Supplementary material. □

We remark that similar results to Theorems 4.2 and 4.3 hold when service preemption and abandonments during service are not allowed.

4.2. Dynamic control

For the multi-server model, we have the following result, which is akin to Proposition 3.1 in [49] on the allocation of servers when the number of customers is sufficiently high. The proof is similar to Proposition 3.1 in [49] and is therefore omitted.

Proposition 4.4. *For the non-collaborative model we are considering, if the number of customers at each queue exceeds the number of servers, there exists a discounted cost optimal control policy that allocates all servers to one phase or the other (i.e., servers are not split between the two queues). Similarly, the result holds in the average cost case when the average cost optimality equations have a solution.*

We remark that if we assume that when more than one server is working at a station, their rates are additive, then, for the multi-server model with abandonments during service and service preemption, there is a discounted cost and average cost optimal policy that does not split the servers between the two phases of service. We also note that Proposition 4.4 does not hold when service preemption is not allowed. Proposition 4.4 implies that when abandonments during service and preemption are allowed, we can restrict attention to policies that always allocate all servers to one station or the other when both $x_1, x_2 \geq N$ (making the service rate $N\mu_1$ or $N\mu_2$). Furthermore, the second part of Proposition 4.1 implies that we should keep as many servers busy as possible in states such that $x_1 + x_2 \leq N$ when service preemption is allowed. However, we were unable to further characterize the optimal policy in states with x_1 or x_2 (but not both) greater than N nor how to choose priorities when $x_1, x_2 \geq N$. In the following, we discuss under what conditions static priority policies are optimal in the single-server model (see Section 4.3) and revisit the multi-server model in our numerical study (Section 5).

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

4.3. The single-server proxy

In this section, we restrict attention to the single-server model ($N = 1$) and provide conditions under which one particular phase of service should be prioritized.

Theorem 4.5. *Under the α-discounted cost criterion, it is optimal to serve at station 2 whenever station 2 is not empty if either one of the following two sets of conditions hold:*

1. $\beta_2 = 0$ *and* $\mu_1[h_1 + \beta_1 K_1 - ph_2] \le \mu_2 h_2$*; or*
2. $\mu_1 = \mu_2$, $\beta_1 - \beta_2 - \mu_2 \ge 0$, *and* $h_1 + \beta_1 K_1 - p(h_2 + \beta_2 K_2) \le h_2 + \beta_2 K_2$.

If, in addition, $\lambda_1 \cdot (1/\pi_0(\mu_1 + \beta_1) + p(1 - P(Ab)/\mu_2) + \lambda_2/\mu_2 < 1$, *conditions 1 or conditions 2 imply that serving station 2 when station 2 is not empty is also average cost optimal.*

Proof of 4.5 when conditions 1 hold. The proof is given for the discounted expected cost model. The proof of the long-run average cost case is similar. Note that the optimality equations imply that it is optimal to prioritize a class 2 customer in state (x_1, x_2) with $x_1, x_2 \ge 1$ when

$$\mu_1[pv_\alpha(x_1 - 1, x_2 + 1) + qv_\alpha(x_1 - 1, x_2) - v_\alpha(x_1, x_2)] + \mu_2[v_\alpha(x_1, x_2) - v_\alpha(x_1, x_2 - 1)] \ge 0. \quad (4.1)$$

We show (4.1) via a sample path argument. Fix $x_1, x_2 \ge 1$ and start five processes on the same probability space. Processes 1–5 begin in states $(x_1 - 1, x_2 + 1)$, $(x_1 - 1, x_2)$, (x_1, x_2), (x_1, x_2), and $(x_1, x_2 - 1)$, respectively. Processes 1, 2, and 4 use stationary optimal policies, which we denote by π_1, π_2, and π_4, respectively. In what follows, we show how to construct (potentially sub-optimal) policies for Processes 3 and 5 which we denote by π_3 and π_5, so that

$$\mu_1[pv_\alpha^{\pi_1}(x_1 - 1, x_2 + 1) + qv_\alpha^{\pi_2}(x_1 - 1, x_2) - v_\alpha^{\pi_3}(x_1, x_2)] + \mu_2[v_\alpha^{\pi_4}(x_1, x_2) - v_\alpha^{\pi_5}(x_1, x_2 - 1)] \ge 0. \quad (4.2)$$

Since π_3 and π_5 are potentially suboptimal, (4.1) follows from (4.2). In what follows, discounting is suppressed without any loss of generality.

Observe that starting from (4.2), the costs incurred until the next event are $(\mu_2 h_2 - \mu_1[h_1 + \beta_1 K_1 - ph_2])t_1 \ge 0$, where t_1 is the time of the next event and the inequality is due to the assumption that $\mu_2 h_2 \ge \mu_1[h_1 + \beta_1 K_1 - ph_2]$. Moreover, if the relative position (as measured by the current states) of the five processes at the next event remains the same, then we may relabel the initial states and continue from the beginning of the argument. This occurs when any of the uncontrolled events occur that are seen by all five processes (i.e., an arrival or an abandonment at station i and $x_i > 1$) . It also occurs when the next event is a service completion and when π_1, π_2, and π_4 serve the same customer class $k \in \{1, 2\}$ by letting π_3 and π_5 also serve the same customer class k customer provided there is one or more customer class k customer in all five processes. Consider now the other cases.

Case 1. Customer abandonments

If $x_1 = 1$ and the first event is a class 1 abandonment in Processes 3–5 only (with probability $\beta_1/(\lambda_1 + \lambda_2 + \mu_1 + \mu_2 + x_1\beta_1)$), after which all processes follow an optimal control, it follows that the remaining costs on the left side of (4.2) (with the probability of this event in the expression suppressed) are

$$\begin{aligned}(\mu_2 h_2 - \mu_1[h_1 + \beta_1 K_1 - ph_2])t_1 + p\mu_1[v_\alpha(x_1 - 1, x_2 + 1) - v_\alpha(x_1 - 1, x_2)] \\ + \mu_2[v_\alpha(x_1 - 1, x_2) - v_\alpha(x_1 - 1, x_2 - 1)]. \end{aligned} \quad (4.3)$$

The terms in this last expression above are nonnegative as a consequence of Proposition 4.1.

There are seven cases left to consider corresponding to service completions all with algebra that is directly analogous. Complete details are available in the Online Appendix in Supplementary material.

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

It follows that, in every case save one (Case 1), we may relabel the states and continue. Thus, to obtain the result, we wait until Case 1 occurs. □

Proof of 4.5 when conditions 2 hold. See Online Appendix in Supplementary material. □

We make several observations about Theorem 4.5 and its associated conditions. First, the policy that prioritizes station 2 is nonpreemptive. Second, if $\min\{\beta_1, \beta_2\} > 0$, then $\mu_1[h_1 + \beta_1 K_1 - p(h_2 + \beta_2 K_2)] \leq \mu_2(h_2 + \beta_2 K_2)$ and $\mu_1[h_1 + \beta_1 K_1 - p(h_2 + \beta_2 K_2)]/\beta_1 \leq \mu_2(h_2 + \beta_2 K_2)/\beta_2$ do not guarantee the optimality of prioritizing station 2, as examples in Section 5 illustrate. Third, the stability condition $\lambda_1 \cdot (p(1 - \Pr(\text{Ab}))/\pi_0\mu_1 + 1/\mu_2) + \lambda_2/\mu_2 < 1$ is derived by computing the fraction of customers arriving to station 1 that enter station 2 and then computing the fraction of time the server works at station 1 (under the prioritize station 2 policy) (see the Online Appendix in Supplementary material for complete details). Fourth, the second condition in 2 says that the station 1 abandonment rate β_1 is higher than the total departure rate from station 2, $\beta_2 + \mu$. If $p = 0$, then the system is the same as the one considered by Down *et al.* [14]. In this case, the term corresponding to $(\beta_2 - \beta_1 - \mu)p$ in subcase 1.2 in the Online Appendix in Supplementary material is 0 so that the same proof of 4.5 when conditions 1 holds, yields that $h_2 + \beta_2 K_2 \geq h_1 + \beta_1 K_1$ and $\beta_1 \geq \beta_2$ imply that it is optimal to prioritize station 2, in agreement with Theorem 3.5 of Down *et al.* [14]. Fifth, one may conjecture that Theorem 4.5 extends to the case when abandonments during service are not allowed, but this turns out not to be so simple since the allocation decision impacts the abandonment rate *and* the current per unit holding cost, suggesting that additional conditions may be required.

Theorem 4.6. *Under the α-discounted cost criterion, it is optimal to serve at station 1 whenever station 1 is not empty if one of the following two sets of conditions hold:*

1. $\beta_1 = 0$ *and* $\mu_2[h_2 + \beta_2 K_2] \leq \mu_1[h_1 - p(h_2 + \beta_2 K_2)]$*; or*
2. $\mu_1 = \mu_2$, $\beta_2 \geq \beta_1$, *and* $h_2 + \beta_2 K_2 \leq h_1 + \beta_1 K_1 - p(h_2 + \beta_2 K_2)$.

If, in addition, $\lambda/\mu_1 < 1$, conditions 1 or 2 imply that serving station 1 when station 1 is not empty is also average cost optimal.

Proof of 4.6. See Online Appendix in Supplementary material. □

We note that when $\min\{\beta_1, \beta_2\} > 0$, $\mu_1[h_1 + \beta_1 K_1 - p(h_2 + \beta_2 K_2)] \geq \mu_2(h_2 + \beta_2 K_2)$ or $\mu_1[h_1 + \beta_1 K_1 - p(h_2 + \beta_2 K_2)]/\beta_1 \geq \mu_2(h_2 + \beta_2 K_2)/\beta_2$ do not guarantee that it is *always* optimal to prioritize station 1 except to avoid unforced idling, as examples in Section 5 illustrate. Furthermore, $\lambda_1/\mu_1 < 1$ is the stability condition for an M/M/1 queue and is sufficient for positive recurrence of all states under the prioritize station 1 policy (see the Online Appendix in Supplementary material for complete details). Lastly, the conditions 2 are the same conditions that guarantee that prioritize station 1 is optimal in Down *et al.* [14], which is a special case of our model when $p = 0$.

5. Numerical study

5.1. Rationale

This section affords us the opportunity to address shortcomings from our analytical analysis. One shortcoming is that our model relies on unrealistic assumptions for systems of interest (i.e., one server, abandonments during service or preemption, and exponential service and abandonment times). Furthermore, the properties of optimal controls remain elusive in many scenarios. In the single-server Markovian model with preemption and abandonments during service, one might conjecture that if it is optimal to serve in station 1 (2) in state (x_1, x_2), then it is also optimal to serve at station 1 (2) when there are more customers in station 1 (2) in state $(x_1 + 1, x_2)$ $((x_1, x_2 + 1))$. To show such a result, standard approaches include sample path arguments or a smoothed rate truncation approach. For either approach, abandonments lead to having to show several inequalities, such as submodularity, of the value function

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

for the discounted cost problem and the relative value function for the long-run average cost problem (cf., [10]). Unfortunately, submodularity remains elusive when the number of jobs at each station is zero; that is, at the boundary of the state space. Despite this, the optimal server allocation policy might still have a threshold or switching curve structure, but proving such a result without submodularity or convexity is difficult. Because the optimal server allocation policy is difficult to discern, we propose and evaluate a class of threshold heuristics that initially assign priority to one phase of service and then switch between priorities based on a state-dependent rule.

Our numerical study has four main goals. The first is to quantify the relative performance of heuristic allocation policies with respect to average costs. The second is to quantify the degree to which (non-priority) threshold policies outperform strict priority policies. The third is to determine how relative performance is affected by changes in the coefficient of variation for the service and abandonment times and changes in

$$\lambda_1 \cdot \left(\frac{1}{\mu_1 + \beta_1} + \frac{p}{\mu_2 + \beta_2} \right) + \frac{\lambda_2}{\mu_2 + \beta_2},$$

which we use as a proxy for the system load. This proxy load corresponds to the average service time for a job that is processed in both phase 1 and then phase 2 service under the prioritize station 2 policy when preemption and abandonments during service are allowed. The fourth is to determine the degree to which guidelines, like the classic c-μ rule, correctly identify the best policy.

5.2. Overview

Our numerical study presents a discrete-event simulation. To simplify exposition, the main text focuses on a multi-server model without preemption and abandonments during service and without exponential service and abandonment times. We vary model parameters to recover guidelines for when one policy might be preferred over another. Simulation results are presented in the Online Appendix in Supplementary material for a single-server Markovian model that allows preemption and abandonments during service and that assumes service and abandonment times are exponentially distributed. The Online Appendix in Supplementary material also benchmarks (non-priority) threshold policies, with respect to long-run average costs, in situations when we know prioritizing a particular phase of service is optimal.

5.3. Heuristic allocation policies

Four types of policies are considered. Each policy is specified to prevent idling, that is, when there are not enough jobs to serve at one phase, providers work at the other phase. When there is enough work to do, we do not assume that providers are split between phases. The first type of policy, denoted by P1(n), prioritizes phase 1 service until there are n total customers in the system after which phase 2 service is prioritized until phase 2 is emptied. When phase 2 is empty, the policy is reset and phase 1 is prioritized again. When $n = \infty$, policy P1(∞) is the priority rule (denoted simply by P1) that always prioritizes phase 1 service over phase 2.

In a reciprocal manner, we define a second type of policy P2(n) to prioritize phase 2 service until there are n total customers in the system after which phase 1 service is prioritized until phase 1 is emptied. When phase 1 is empty, the policy is reset and phase 2 is prioritized again. When $n = \infty$, policy P2(n) is simply the priority rule for phase 2 (denoted by P2). Another policy considered, denoted by Exh, is an exhaustive policy. It prioritizes one phase of service until that phase is empty after which it switches to the other phase until that phase is empty. It continues to switch between phases, emptying each phase in turn before switching to the other. The last policy considered, denoted by Inc, prioritizes whichever phase of service has more customers. It is considered an increasing policy, because the threshold of phase 2 customers for switching as a function of phase 1 customers is an increasing function.

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

Table 1. *Parameters used for the simulation.*

Parameters	Description	Range
λ_1	Arrival rate at 1	9
λ_2	Arrival rate at 2	[0,3]
μ_1	Service rate at 1	[4,12]
μ_2	Service rate at 2	[4,12]
β_1	Abandonment rate at 1	[0.1,3]
β_2	Abandonment rate at 2	[0.1,3]
p	Joining probability	[0.25,1]
h_1	Holding cost rate at 1	[0.1,3]
h_2	Holding cost rate at 2	[0.1,3]
K_1	Abandonment cost at 1	[0.1,3]
K_2	Abandonment cost at 2	1
cv	Coefficient of variation	[0.6,1.4]
nw	Number of workers	3

Since P1(n) and P2(n) prioritize different phases of service, we explore the degree to which one policy outperforms another when the inequality

$$\mu_1 h_1 \leq \mu_2 h_2 \tag{5.1}$$

holds as well as when

$$\mu_1(h_1 + \beta_1 K_1 - p(h_2 + \beta_2 K_2)) \leq \mu_2(h_2 + K_2\beta_2) \tag{5.2}$$

holds. The former inequality reflects the classic c-μ rule for deciding when to prioritize one phase of service over another when there are no abandonments and the latter inequality reflects the analog of the c-μ rule when you account for abandonments and are part of the sufficient conditions we provided to guarantee the optimality of a policy that prioritizes phase 1 (policy P1) or phase 2 (policy P2).

5.4. Parameters

Parameters are summarized in Table 1 for the simulation. Parameters were chosen to capture situations when the optimal policy remains elusive. Given the importance of the classic c-μ inequality (Eq. (5.1)) and its extended version (Eq. (5.2)), parameters were selected and varied to both satisfy and violate these situations. It is without any loss of generality that we can fix one cost and one rate. So, the abandonment cost K_2 at phase 2 is fixed at 1 and the arrival rate λ_1 at phase 1 is fixed at 9. Assuming a time unit of hours, we simulated the system for each parameter set over a simulated time horizon of 5 years after a 5 year warm-up period and then performed 50 replications of this simulation. Average costs were averaged over the time horizon and then over the replications.

Parameters have a similar interpretation as the single-server Markovian model, with the following exceptions. First, we fixed the number of workers to be 3. Second, abandonment and service times were modeled as Gamma random variables as opposed to exponential random variables. Gamma shape parameters ranged from $1/2$ or 3, yielding random times that have standard deviations larger than their mean and smaller than their mean, complementing exponential random times, which have standard deviations equal to their mean. Coefficient of variations (cv) were respectively 1.4 and 0.6 for the two shape values. Last, parameters μ_1, μ_2, β_1, and β_2 refer to *average* rates, which meant that the rate parameters for the gamma distributions needed to be μ_1, μ_2, β_1, β_2 scaled by the corresponding shape parameter.

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

Table 2. *Percent samples for which given policy yields lowest average costs as a function of service rates (μ_1 and μ_2) and approximate load under P2.*

			Policy							
μ_1	μ_2	≈P2 load	P1	P2	P1(5)	P2(5)	Exh	Inc	c-μ	Ext. c-μ
All	All	2.2	19.2	64.1	5.7	5.2	0.8	5.0	59.1	65.2
4	4	3.1	17.9	56.7	7.6	10.3	0.9	6.7	47.7	43.9
4	12	2.3	1.7	86.8	2.8	1.3	1.5	5.8	80.2	79.1
12	4	2	38.3	41.6	5.9	7.7	0.5	6	47.6	59.2
12	12	1.2	18.9	71.1	6.6	1.4	0.3	1.7	61	78.4

We first explored parameter space using a full factorial design of six parameters (λ_2, μ_1, μ_2, β_1, β_2, p, and cv); each parameter had two levels corresponding to the lowest and highest value in the parameter range listed in Table 1. For each of these 128 sets of parameters, we then sampled 10,000 sets of costs (h_1, h_2, K_1) uniformly from the parameter range listed in Table 1. We then systematically varied parameters while keeping fixed (unless otherwise specified) $\lambda_1 = 9$, $\lambda_2 = 0$, $\mu_1 = \mu_2 = 8$, $p = \beta_1 = \beta_2 = h_1 = h_2 = K_2 = 1$, and $K_1 = 2$. Service rates μ_1 and μ_2 were systematically varied, followed by abandonment rates β_1 and β_2, holding cost rates h_1 and h_2, and arrival rate λ_2 and joining probability p.

5.5. Results

We first report relative performance of the policies over 1,280,000 samples of parameter space for the multi-server model (Table 2). Policy P2 had lowest average costs for 64.1% of these samples compared to 19.2% of samples for P1, 5.7% for P1(5), 5.2% for P2(5), 5.0% for Inc, and less than 1% for Exh. Following P2 is even better than using the classic c-μ inequality (Eq. (5.1)) to guide when to use P2 over P1, the latter being best for only 59.1% of samples compared to the aforementioned 64.1% for P2. Following P2, however, is slightly worse than using the extended version of the c-μ inequality (Eq. (5.2)) to guide when to use P2 over P2, which is best for 65.2% of samples.

Samples were stratified by service rates μ_2 and μ_1 (Table 2). When μ_2 is high and μ_1 is low, policy P2 is, as expected, best for 86.8% of samples and for more samples than following the c-μ rule or its extended version. Policy P2 is best for fewer samples (56.7%) when both μ_1 and μ_2 are low. In this case, P2 is again better for more samples than the c-μ rule or its extended version. Meanwhile, policy P2 is best for only 41.6% of samples and for fewer samples than the extended c-μ rule (59.2%) when μ_1 is high and μ_2 is low. When μ_1 and μ_2 are both high, policy P2 is once again best for a majority of samples (71.1%) but still best for fewer samples than the extended c-μ rule (78.4%). Importantly, the extended c-μ rule performs better than P2 for lower approximate loads under P2 (averaged over the relevant samples).

Figure 1 depicts an example when the extended c-μ rule helps guide whether to use P2 or P1. Parameters μ_1 and μ_2 were systematically varied. Other parameters were fixed at values specified earlier; the coefficient of variation (cv) was 1.4. Policy P2 is best for most μ_1 and μ_2 values except when the extended c-μ inequality favors P1 or is close to favoring P1. Policy P1 is best for remaining μ_1 and μ_2 values except in one case when policy Inc is best. Other figures are presented in Appendices A.1 and A.2.

Samples were then stratified by service rates and the cv for service and abandonment times (Table 3). For each pair of service rates μ_1 and μ_2, the policy P2 is the best policy for more samples when the cv is high (1.4) versus low (0.6). This improvement in P2 comes at the expense of P1, wherein P1 is the best policy for an increasing number of samples when the cv increases from 0.6 to 1.4. Even with the additional stratification on cv, the extended c-μ rule still performs better than P2 for lower approximate loads at P2.

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

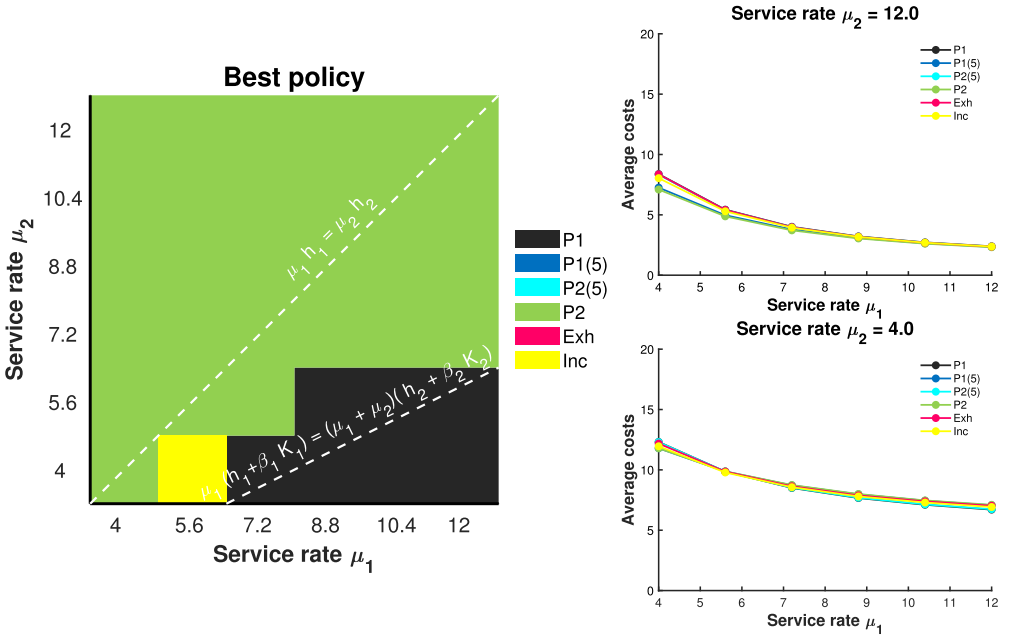

Figure 1. *Average cost comparison for the multi-server model when the cv is fixed at* 1.4 *and service rates* μ_1 *and* μ_2 *are varied.*

Table 3. *Percent samples for which given policy yields lowest average costs as a function of service rates* (μ_1 *and* μ_2)*, approximate load under P2, and coefficient of variation (cv).*

				Policy							
μ_1	μ_2	cv	≈P2 load	P1	P2	P1(5)	P2(5)	Exh	Inc	c-μ	Ext. c-μ
4	4	0.6	3.1	25.2	52.7	9.4	10.2	0.3	2.3	49	40.4
4	4	1.4	3.1	10.5	60.7	5.8	10.4	1.5	11.1	46.5	47.4
4	12	0.6	2.3	2.8	75.5	5.6	2.2	2.4	11.5	73.7	71.4
4	12	1.4	2.3	0.6	98.2	0	0.5	0.5	0.1	86.6	86.8
12	4	0.6	2	43.5	37.1	7.4	6.6	1	4.4	52.3	57.7
12	4	1.4	2	33.1	46.1	4.4	8.7	0.1	7.6	43	60.7
12	12	0.6	1.2	23.7	58.4	12.3	1.9	0.3	3.3	66.8	72.3
12	12	1.4	1.2	14.1	83.8	0.9	0.8	0.3	0	55.2	84.4

With P2 performing best in most samples, we wanted to characterize parameter values when a policy other than P2 performs well. We start with P1. Among the 128 cases of parameters in our factorial design (μ_1, μ_2, β_1, β_2, p, and cv), P1 is best for a majority of samples for 20 of these cases compared to 88 for P2 (Table A.1 in Appendix). As one would expect, all 20 cases have a service rate μ_1 faster or equal to μ_2. Moreover, at least 16 of these cases have at least a high value μ_1, low value of μ_2; or low joining probability p. For 15 of these 20 cases, the extended c-μ rule performs better than P1; these 15 cases coincide exactly with an approximate load under P2 lower than its mean of 2.2. In addition to a larger load, all of the remaining five cases are accompanied with low μ_2, low β_1, high β_2, and low cv; neither λ_2, μ_1, or p took a consistent value across these cases.

Table 4 identifies parameter values (μ_1, μ_2, β_1, β_2, p, and cv) when *(non-priority) threshold policies* are best for a majority of samples. We make three observations. First, there is no set of parameters,

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

Table 4. *Percent samples that policy yields lowest average costs in parameter cases when (non-priority) threshold policies are best for a majority of samples.*

								Policy							
λ_2	μ_1	μ_2	β_1	β_2	p	cv	≈P2 load	P1	P2	P1(5)	P2(5)	Exh	Inc	c-μ	Ext. c-μ
P1(5) is best for majority of samples															
0	4	4	0.1	0.1	1	0.6	4.4	0	0	70.8	29.2	0	0	0	0
0	12	4	0.1	0.1	1	0.6	2.9	12.3	9.3	51.6	26.8	0	0	21.6	21.6
P2(5) is best for majority of samples															
3	4	4	0.1	0.1	0.25	0.6	3.5	0	43.9	0	56.1	0	0	43.9	43.9
3	4	4	0.1	3	0.25	1.4	2.9	1.0	14.9	4.4	55.6	0	24.1	13.9	14.9
3	4	4	0.1	3	1	1.4	3.9	3.4	1.5	11.8	55.5	0	27.9	4.3	1.5
Inc is best for majority of samples															
0	4	12	0.1	3	1	0.6	2.8	2.7	23.4	0	0.3	0	73.6	25.5	23.4
3	4	12	0.1	3	1	0.6	3.0	3.1	12.8	0	0	28.7	55.4	15.7	12.8

among the 128 considered, in which policy Exh performs best for a majority of samples. Second, the abandonment rate β_1 at phase 1 is low for every parameter case when a non-strict priority rule is optimal; neither λ_2, μ_1, μ_2, β_2, p, or cv took a consistent value across these cases. Third, the approximate load under P2 was larger than its average of 2.2 in all these cases when a (non-priority) threshold policy is optimal.

We summarize insights gained from the numerical study as follows. Using the extended c-μ rule should guide when to use P2 or P1 policies, provided the approximate load under P2 is low. When the approximate load under P2 is high, then P2 is likely the best policy, but not always. The poorer relative performance of P2 and of the extended c-μ rule occurs when customers abandon from phase 1 at a slow rate β_1, which can lead to long queues at phase 1, or when the variability in abandonment and service times is smaller than the mean time. Non-strict priority rules such as Inc, P1(5), and P2(5) (though not Exh) may even perform better than P1 or P2 in these situations.

6. Conclusion

In this paper, we allow for abandonments to a stochastic scheduling model consisting of a two-class, two stage tandem service system where we have generalized the models considered in Down *et al.* [14] and Zayas-Cabán *et al.* [49] by allowing customers to join phase 2 service after completing phase 1 service and by considering the performance criterion of minimizing holding costs per customer per unit time and lump-sum abandonment costs. Our main results are conditions for the optimality of static priority rules for the single-server model. Abandonments lead to technical challenges since the abandonment rate is not bounded, and uniformization is not possible. This means the standard induction arguments cannot be readily used. Furthermore, interchange arguments are difficult to apply since customers may abandon in between services. We use the CTMDP framework to analyze this decision-making scenario, and in particular, use the continuous-time optimality equations, and a sample path argument to show the results.

Because the optimal policy remains elusive outside the conditions provided in Theorems 4.5 and 4.6 and because our model relies on assumptions that may be unrealistic for certain systems of interest, we compare the long-run average costs between priority rules, based on the direct analog of the classic c-μ rule when we add abandonments, with more complicated threshold policies in a discrete-event simulation study. We focus on a multi-server model where preemption and abandonments during service are not allowed, and when service and abandonment times are no longer exponentially distributed.

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

There are several venues for future research. Characterizing the optimal policy in general is of clear interest. For example, one way to show the optimality of switching curve policies is to show convexity and submodularity of the value function for the discounted cost problem and of the relative value function for the long-run average cost problem. We have attempted to prove these results in general, but up to this point have been unable to do so. Extending versions of Theorems 4.5 and 4.6 to when abandonments during service are not allowed is also of interest. In the latter case, the allocation decision now impacts the abandonment rates and holding costs per unit time, which suggest stronger conditions than those in Theorems 4.5 and 4.6 are needed to prove the optimality of static priority rules. Another extension is one in which there multiple classes in each phase of service, but one class always having priority over the other. While priorities reduce the complexity of such a proof using inductive or sample path arguments, the number of cases to consider remains high. Lastly, another venue is the study of asymptotically optimal policies akin to the analysis in Larrañaga *et al.* [31] or in James *et al.* [19] for parallel queues with abandonments.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/S0269964822000213.

References

[1] Ahn, H.-S. & Righter, R. (2006). Dynamic load balancing with flexible workers. *Advances in Applied Probability* 38(3): 621–642.

[2] Ahn, H.-S., Duenyas, I., & Zhang, R.Q. (1999). Optimal stochastic scheduling of a two-stage tandem queue with parallel servers. *Advances in Applied Probability* 31(4): 1095–1117.

[3] Ahn, H.-S., Duenyas, I., & Lewis, M.E. (2002). Optimal control of a two-stage tandem queuing system with flexible servers. *Probability in the Engineering and Informational Sciences* 16(4): 453–469.

[4] Ahn, H.-S., Duenyas, I., & Zhang, R.Q. (2004). Optimal control of a flexible server. *Advances in Applied Probability* 36(1): 139–170.

[5] Andradóttir, S. & Ayhan, H. (2005). Throughput maximization for tandem lines with two stations and flexible servers. *Operations Research* 53(3): 516–531.

[6] Andradóttir, S., Ayhan, H., & Down, D.G. (2001). Server assignment policies for maximizing the steady-state throughput of finite queueing systems. *Management Science* 47(10): 1421–1439.

[7] Andradóttir, S., Ayhan, H., & Down, D.G. (2003). Dynamic server allocation for queueing networks with flexible servers. *Operations Research* 51(6): 952–968.

[8] Argon, N.T. & Tsai, Y.-C. (2012). Dynamic control of a flexible server in an assembly-type queue with setup costs. *Queueing Systems* 70(3): 233–268.

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

[10] Bhulai, S., Brooms, A.C., & Spieksma, F.M. (2014). On structural properties of the value function for an unbounded jump Markov process with an application to a processor sharing retrial queue. *Queueing Systems* 76: 425–446.

[11] Bhulai, S., Blok, H., & Spieksma, F.M. (2019). K competing queues with customer abandonment: Optimality of a generalised $c\mu$-rule by the smoothed rate truncation method. *Annals of Operations Research*: 1–30. https://doi.org/10.1007/s10479-019-03131-3

[12] Blok, H. & Spieksma, F. (2015). Countable state Markov decision processes with unbounded jump rates and discounted cost: Optimality equation and approximations. *Advances in Applied Probability* 47(4): 1088–1107.

[13] Blok, H. & Spieksma, F.M. (2017). Structures of optimal policies in MDPs with unbounded jumps: The state of our art. In Boucherie, R. & van Dijk, N. (eds), *Markov decision processes in practice*. International Series in Operations Research & Management Science, vol 248. Cham: Springer, pp. 131–186.

[14] Down, D.G., Koole, G., & Lewis, M.E. (2011). Dynamic control of a single-server system with abandonments. *Queueing Systems* 67(1): 63–90.

[15] Duenyas, I., Gupta, D., & Olsen, T.L. (1998). Control of a single-server tandem queueing system with setups. *Operations Research* 46(2): 218–230.

[16] Guo, X. & Hernández-Lerma, O. (2009). Continuous-time Markov decision processes. In *Continuous-time Markov decision processes*. Stochastic Modelling and Applied Probability, vol 62. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-02547-1_2.

[17] Iravani, S., Posner, M.J.M., & Buzacott, J.A. (1997). A two-stage tandem queue attended by a moving server with holding and switching costs. *Queueing Systems* 26(3–4): 203–228.

[18] Işık, T., Andradóttir, S., & Ayhan, H. (2016). Optimal control of queueing systems with non-collaborating servers. *Queueing Systems* 84(1–2): 79–110.

[19] James, T., Glazebrook, K., & Lin, K. (2016). Developing effective service policies for multiclass queues with abandonment: Asymptotic optimality and approximate policy improvement. *INFORMS Journal on Computing* 28(2): 251–264.
[20] Johri, P. & Katehakis, M.N. (1988). Scheduling service in tandem queues attended by a single-server. *Stochastic Analysis and Applications* 6(3): 279–288.
[21] Kamali, M.F., Tezcan, T., & Yildiz, O. (2018). When to use provider triage in emergency departments. *Management Science* 65(3): 1003–1019.
[22] Katayama, T. (1980). Analysis of an exhaustive service type tandem queue attended by a moving server with walking time. *Transactions of IECE* J63-B: 1055–1062.
[23] Katayama, T. (1981). Analysis of a tandem queueing system with gate attended by a moving server with walking time. *Transactions of IECE* J64-B: 931–938.
[24] Katayama, T. (1983). Analysis of a finite limiting service tandem queue attended by a moving server with walking time. *Review of the Electrical Communication Laboratory* 31: 439–446.
[25] Katayama, T. (1992). Performance analysis and optimization of a cyclic-service tandem queueing system with multi-class customers. *Computers & Mathematics with Applications* 24(1–2): 25–33.
[26] Katayama, T. & Kobayashi, K. (1995). Sojourn time analysis for a cyclic-service tandem queueing model with general decrementing service. *Mathematical and Computer Modelling* 22(10–12): 131–139.
[27] Kaufman, D.L., Ahn, H.-S., & Lewis, M.E. (2005). On the introduction of an agile, temporary workforce into a tandem queueing system. *Queueing Systems* 51(1–2): 135–171.
[28] Koçağa, Y.L. & Ward, A.R. (2010). Admission control for a multi-server queue with abandonment. *Queueing Systems* 65(3): 275–323.
[29] Konrad, R., DeSotto, K., Grocela, A., McAuley, P., Wang, J., Lyons, J., & Bruin, M. (2013). Modeling the impact of changing patient flow processes in an emergency department: Insights from a computer simulation study. *Operations Research for Health Care* 2(4): 66–74.
[30] Larranaga, M., Ayesta, U., & Verloop, I.M. (2013). Dynamic fluid-based scheduling in a multi-class abandonment queue. *Performance Evaluation* 70(10): 841–858.
[31] Larrañaga, M., Ayesta, U., & Verloop, I.M. (2015). Asymptotically optimal index policies for an abandonment queue with convex holding cost. *Queueing Systems* 81(2–3): 99–169.
[32] Legros, B., Jouini, O., & Koole, G. (2017). A uniformization approach for the dynamic control of queueing systems with abandonments. *Operations Research* 66(1): 200–209.
[33] Liang, H.M. & Kulkarni, V.G. (1999). Optimal routing control in retrial queues. In Shanthikumar, J.G. & Sumita, U. (eds), *Applied probability and stochastic processes*. International Series in Operations Research & Management Science, vol 19. Boston, MA: Springer, pp. 203–218.
[34] Nair, S. (1971). A single server tandem queue. *Journal of Applied Probability* 8(1): 95–109.
[35] Nelson, R. (1966). Labor assignment as a dynamic control problem. *Operations Research* 14(3): 369–376.
[36] Pandelis, D.G. (2007). Optimal use of excess capacity in two interconnected queues. *Mathematical Methods of Operations Research* 65(1): 179–192.
[37] Pandelis, D.G. (2008). Optimal control of flexible servers in two tandem queues with operating costs. *Probability in the Engineering and Informational Sciences* 22(1): 107–131.
[38] Prieto-Rumeau, T. & Hernández-Lerma, O. (2012). *Selected topics on continuous-time controlled Markov chains and Markov games*, vol. 5. London: World Scientific.
[39] Rastpour, A., Ingolfsson, A., & Sandıkçı, B. (2020). Algorithms for queueing systems with reneging and priorities modeled as quasi-birth-death processes. Technical Report, Working Paper, University of Ontario Institute of Technology, Ontario, Canada.
[40] Saghafian, S., Hopp, W.J., Van Oyen, M.P., Desmond, J.S., & Kronick, S.L. (2014). Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing and Service Operations Management* 16(3): 329–345.
[41] Schiefermayr, K. & Weichbold, J. (2005). A complete solution for the optimal stochastic scheduling of a two-stage tandem queue with two flexible servers. *Journal of Applied Probability* 42(3): 778–796.
[42] Sidi, M., Levy, H., & Fuhrmann, S.W. (1992). A queueing network with a single cyclically roving server. *Queueing Systems* 11(1–2): 121–144.
[43] Taube-Netto, M. (1977). Two queues in tandem attended by a single server. *Operations Research* 25(1): 140–147.
[44] van Dijk, N.M. (1988). On the finite horizon bellman equation for controlled Markov jump models with unbounded characteristics: Existence and approximation. *Stochastic Processes and Their Applications* 28(1): 141–157.
[45] Van dijk, N.M. (1989). A note on constructing e-optimal policies for controlled Markov jump models with unbounded characteristics. *Stochastics: An International Journal of Probability and Stochastic Processes* 27(1): 51–58.
[46] Van Oyen, M.P., Gel, E.G., & Hopp, W.J. (2001). Performance opportunity for workforce agility in collaborative and noncollaborative work systems. *IIE Transactions* 33(9): 761–777.
[47] Wang, J., Abouee-Mehrizi, H., Baron, O., & Berman, O. (2019). Tandem queues with impatient customers. *Performance Evaluation* 135: 102011.
[48] Zayas-Cabán, G. & Ahn, H.-S. (2018). Dynamic control of a single-server system when jobs change status. *Probability in the Engineering and Informational Sciences* 32(3): 353–395.

[49] Zayas-Cabán, G., Xie, J., Green, L.V., & Lewis, M.E. (2016). Dynamic control of a tandem system with abandonments. *Queueing Systems* 84(3–4): 279–293.
[50] Zayas-Caban, G., Xie, J., Green, L.V., & Lewis, M.E. (2019). Policies for physician allocation to triage and treatment in emergency departments. *IISE Transactions on Healthcare Systems Engineering* 9(4): 342–356.

A. Appendix

Parameter cases when P1 performs best

A.1. Additional simulations when cv = 1.6

Parameters were systematically varied to examine further when to use various policies, keeping other parameters fixed. The coefficient of variation, cv, was fixed at 1.6. Recall that the main text reports the case when μ_1 and μ_2 are systematically varied. Varying holding cost rates h_1 and h_2 also shows that P2 performs well when the extended c-μ inequality favors P2 (Figure A.1). Policy P1 performs well when this inequality favors P1. In addition, we again find cases when a non-priority rule, in this case policy Exh, outperforms both priority rules, but the gap between Exh and P2 is small in these cases (~0.1%).

Varying abandonment rates β_1 and β_2 (Figure A.2) yields situations when policy P2 can perform better than P1 even when the extended c-μ inequality favors P1. In addition, we find several cases when the threshold policy P1(5) performs better than the other heuristic policies, albeit it is close to P2 (~0.3% away). By contrast, the extended c-μ inequality provides perfect guidance when varying the arrival rate λ_2 to phase 2 or the joining probability p (Figure A.3). That is, for these values of p and λ_2, policy P2 performs better than P1 when the extended c-μ inequality favors P2, whereas policy P1 performs better than P2 when the extended c-μ inequality favors P1.

A.2. Additional simulations when cv = 0.4

Parameters were again systematically varied as before except the coefficient of variation, cv, is set to 0.4 as opposed to 1.6. Here, policy P2 performs best for all choices of μ_1 and μ_2, but its improvement over

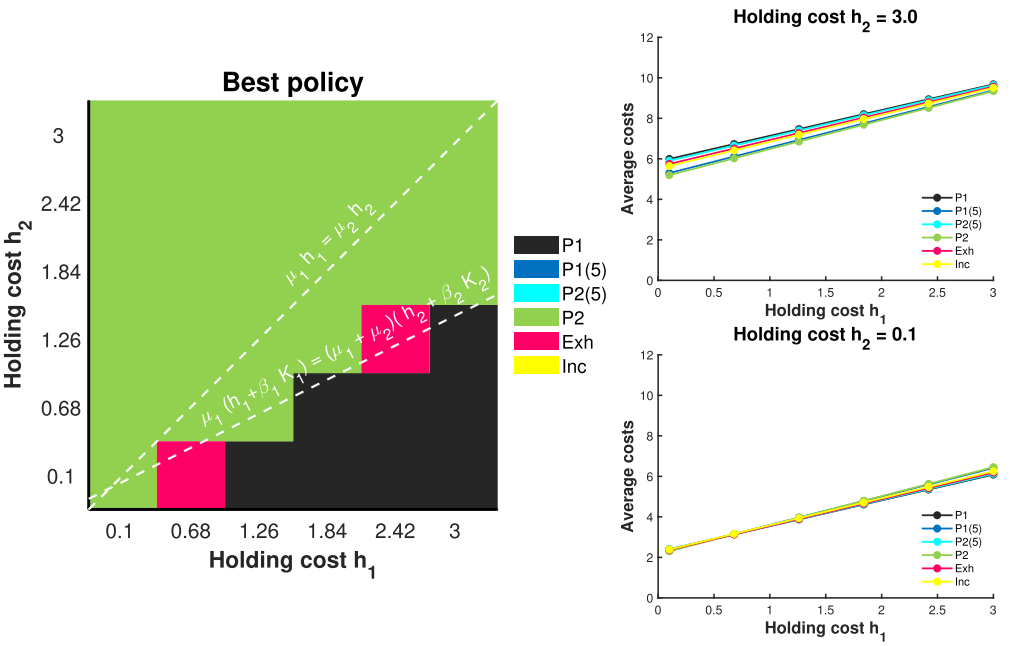

Figure A.1. *Average cost comparison for the multi-server model when cv is 1.6 and holding cost rates h_1 and h_2 are varied.*

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

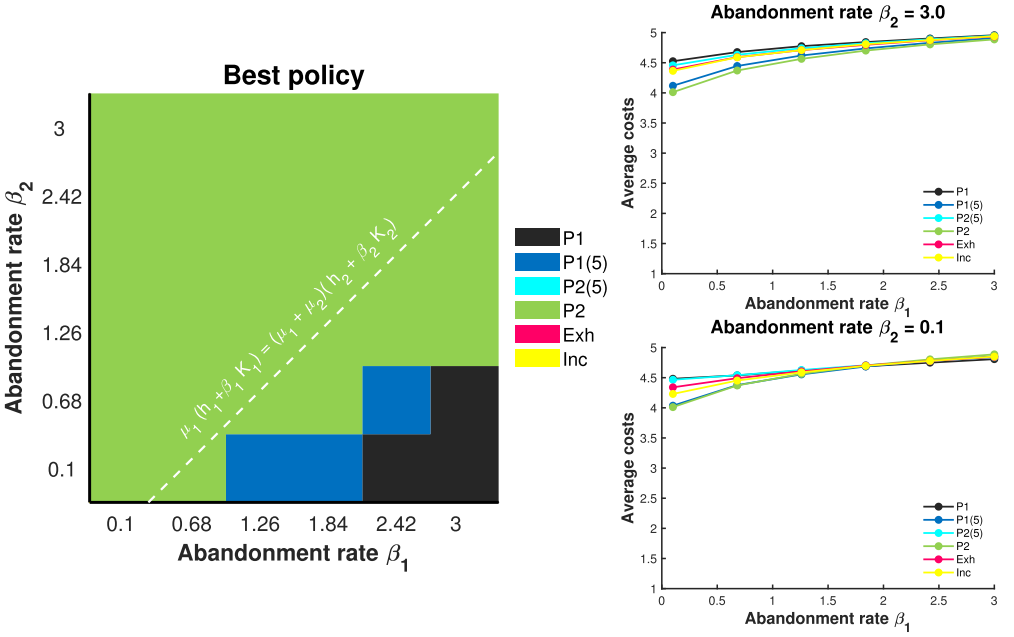

Figure A.2. *Average cost comparison for the multi-server model when cv is 1.6 and abandonment rates β_1 and β_2 are varied.*

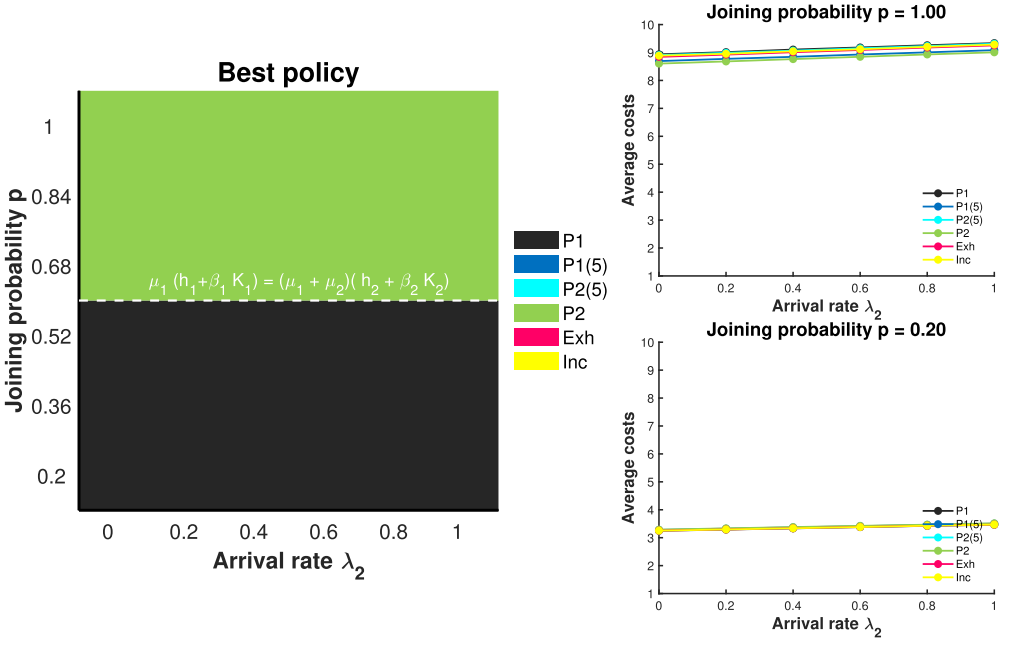

Figure A.3. *Average cost comparison for the multi-server model when cv is 1.6 and the joining probability p and arrival rate λ_2 are varied.*

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

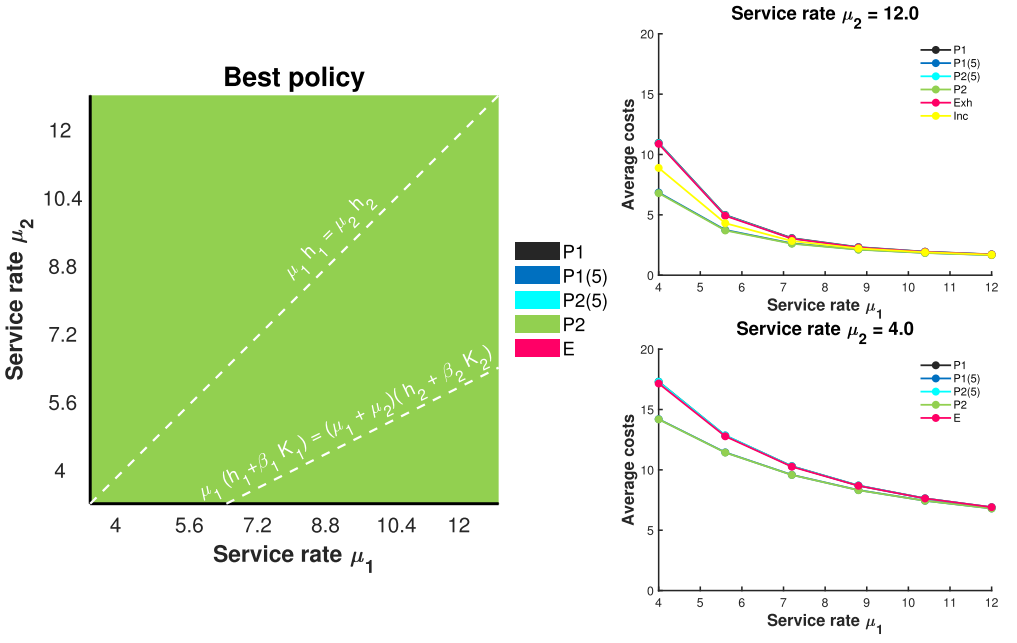

Figure A.4. *Average cost comparison for the multi-server model when the cv is 0.4 and service rates μ_1 and μ_2 are varied.*

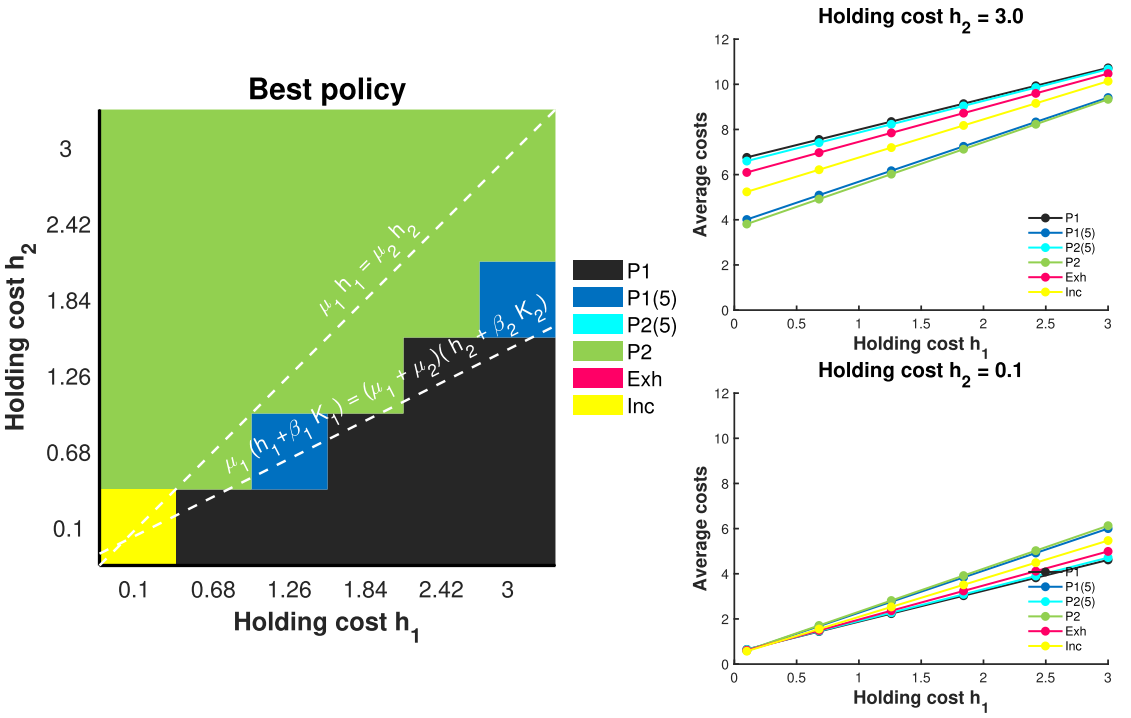

Figure A.5. *Average cost comparison for the multi-server model when the cv is 0.4, and holding cost rates h_1 and h_2 are varied.*

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

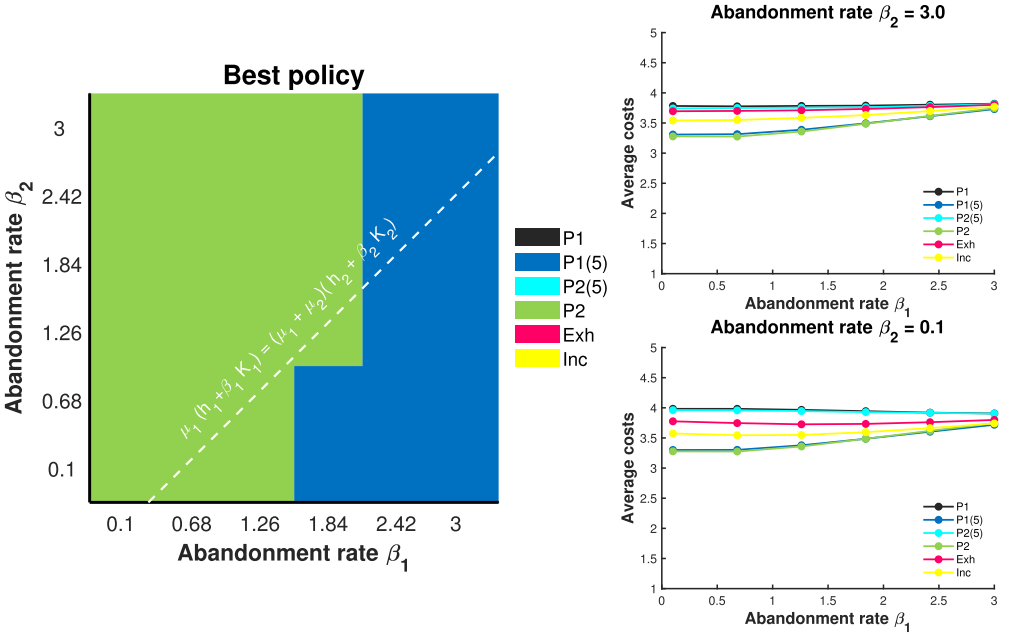

Figure A.6. *Average cost comparison for the multi-server model when the cv is 0.4, and abandonment rates β_1 and β_2 are varied.*

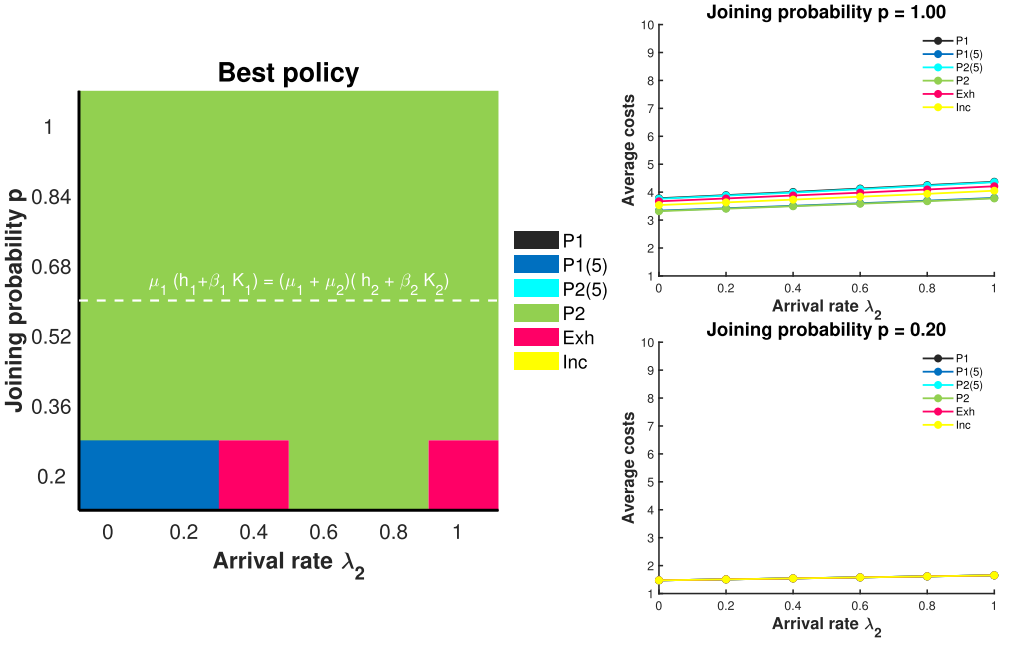

Figure A.7. *Average cost comparison for the multi-server model when the cv is 0.4 and joining probability p and arrival rate λ_2 are varied.*

P1 is small (<5%) when the extended c-μ inequality is violated. Thus, the extended c-μ inequality can still help guide which priority rule performs well when varying service rates μ_1 and μ_2, but may not always yield the best policy (Figure A.4).

https://doi.org/10.1017/S0269964822000213 Published online by Cambridge University Press

Table A.1. *Percent samples that policy yields lowest average costs in parameter cases when P1 is best for a majority of samples.*

								Policy							
λ_2	μ_1	μ_2	β_1	β_2	p	cv	≈P2 load	P1	P2	P1(5)	P2(5)	Exh	Inc	c-μ	Ext. c-μ
0	4	4	0.1	3.0	1.00	0.6	3.5	93.8	0	3.5	0.4	1.9	0.4	49.1	0
3	4	4	0.1	3.0	0.25	0.6	2.9	86	0.3	0	0	0	13.7	48.3	0.3
0	12	4	3.0	0.1	0.25	0.6	1.1	85	10.4	0.7	3.8	0	0.2	82	86.4
3	4	4	0.1	3.0	1.00	0.6	3.9	81.6	0	4.3	14.2	0	0	49.1	0
0	12	4	3.0	0.1	0.25	1.4	1.1	77.9	18.7	3.4	0.1	0	0	73	79.3
3	12	4	3.0	0.1	0.25	0.6	1.9	76	19.9	0.5	0	0.6	2.9	74.4	77.4
3	12	4	3.0	0.1	0.25	1.4	1.9	72.2	17.7	7.9	0.5	0	1.8	68.5	73.5
0	12	4	3.0	3.0	0.25	0.6	0.9	69	17.8	4.9	4	0	4.3	70.7	83.3
3	12	4	3.0	3.0	0.25	0.6	1.4	68.8	23.9	4	1.7	0	1.7	69.2	91.9
3	12	12	3.0	0.1	0.25	1.4	1.0	66.8	29.3	3	0.1	0	0.7	56.9	75.3
0	12	4	0.1	0.1	0.25	0.6	1.3	66.8	30.6	0	2.7	0	0	81.4	93.7
3	12	4	0.1	3.0	1.00	0.6	2.5	66.5	0.4	0.5	23.4	1.5	7.7	66.9	0.4
0	12	12	3.0	0.1	0.25	1.4	0.8	66.2	33.3	0.5	0	0	0	58.6	74.7
3	12	4	0.1	0.1	0.25	1.4	2.0	63.5	26.2	0	8.5	0	1.8	77.3	86.1
3	12	4	0.1	0.1	0.25	0.6	2.0	63.4	23.4	5.1	7.4	0.8	0	78	86.5
3	12	12	3.0	0.1	0.25	0.6	1.0	60.8	29.5	0	9.6	0	0.1	74.6	69.2
0	12	4	0.1	0.1	0.25	1.4	1.3	58.9	33.8	0	7.2	0	0	73.3	84.7
3	12	4	3.0	3.0	0.25	1.4	1.4	57.3	30.5	9.8	2.5	0	0	55	83.1
0	12	12	3.0	0.1	0.25	0.6	0.8	57.1	25.6	0.7	13.4	3.1	0	70.2	65.5
0	4	4	0.1	3.0	0.25	0.6	2.5	56.4	9.9	0	29.4	0	4.3	52.1	9.9

Meanwhile, varying holding cost rates h_1 and h_2 (Figure A.5) reinforces the use of the extended c-μ inequality. Policy P2 performs well when this inequality is satisfied and policy P2 performs well when this inequality is not satisfied. In addition, we again find cases when a threshold policy, that is, P1(5), that outperforms both priority rules, but the gap between this threshold policy and P2 is small in these cases (~0–1%).

Varying abandonment rates β_1 and β_2 (Figure A.6) yields situations when policy P2 can perform better than P1 even when the extended c-μ inequality is violated. This occurs when the abandonment rate β_2 is low, reinforcing what we found in Scenario 1 (and in the single-server model): that neglecting phase 2 when there are few abandonments at 2 can yield poor performance. In addition, we find several cases when the threshold policy P1(5) performs better than the other heuristic policies, albeit it is close to P2 ($\leq$1% away).

Last, we varied the joining probability p and the arrival rate λ_2 (Figure A.7). Similar to what we observed when we varied μ_1 and μ_2, policy P2 performs better than policy P1 in all choices of p and λ_2, and best in most in most choices, but its improvement over P1 is small (<1%) when the extended c-μ inequality is violated. Thus, the extended c-μ inequality can still help guide which priority rule performs well when varying service rates μ_1 and μ_2, but may not always yield the best policy.

Cite this article: Zayas-Cabán G and Cochran AL (2023). Scheduling servers in a two-stage queue with abandonments and costs. *Probability in the Engineering and Informational Sciences* **37**, 833–851. https://doi.org/10.1017/S0269964822000213